وزارة التعليم العالي و البحث العلمي

Université Badji Mokhtar –Annaba-

جامعة باجي مختار - عنابة-

Badji Mokhtar –Annaba- University

Année: **2015**

Faculté des sciences de l'ingénieur
Département d'informatique

THÈSE

Pour obtenir le diplôme de
Docteur 3^{éme} cycle

La Sémantique et l'Effet Communautaire: Enrichissement et Exploitation

Filière : Informatique
Spécialité : Sciences et Technologies de l'Information et de la Communication

Préparée par

Samia Beldjoudi

Jury :

Présidente: Mme Labiba Souici	Pr. Université Badji Mokhtar Annaba
Directrice de thèse: Mme Hassina Seridi	Pr. Université Badji Mokhtar Annaba
Co-directrice de thèse: Mme Catherine Faron-Zucker	Pr. Université Sophia Antipolis Nice (France)
Examinateur : Mr Nadir Farah	Pr. Université Badji Mokhtar Annaba
Examinatrice : Mme Halima Bahi	Pr. Université Badji Mokhtar Annaba
Examinatrice : Mme Amel Yessad	Dr. Université Pierre et Marie Curie Paris (France)

Acknowledgements

First of all, I would like to thank my supervisors Professors Hassina Seridi and Catherine Faron-Zucker for their excellent supervision and great support during the 5 years of my PhD study. Not only teaching me how to conduct research, they have also given me an insight into my future career. I have enjoyed a lot doing research because of my supervisors and a highly motivating research topic. I am grateful to their support, patience, and all the many things I learned from both of them.

My sincere thanks to Professor Labiba Souici for agreeing to chair the evaluation jury. And I thank also Professors Nadir Farah, Halima Bahi and Doctor Amel Yessad for agreeing to be part of the examination board.

I thank my family especially my parents, my sister, my brothers and especially my husband for their support.

Finally, I extend my sincere gratitude and appreciation to my friends and all people who made this PhD thesis possible.

Abstract

Collaborative tagging which is the keystone of social practices in web 2.0 has been highly developed in the last few years. The powerful of social tagging activities allows the wide set of web users to add free annotations on resources to express user interests, preferences and also automatically generate folksonomies. Folksonomies have been involved in many web searching and recommendations approaches, consequently, it seems logical to use information about and derived from social network structures in the context of information retrieval systems. The presented PhD thesis addresses different challenges in the social web area. The main focus of our dissertation is how to exploit social aspect of folksonomies within information retrieval, especially in searching and recommendation.

We proposed a new method to analyzing user profiles according to their tagging activity in order to improve resource recommendation. We based upon association rules which are a powerful method to discover interesting relationships among large datasets on the web. Our aim is to recommend resources annotated with tags suggested by association rules, in order to enrich user profiles. The effectiveness of recommendation depends on the resolution of social tagging drawbacks. In our recommender process, we demonstrate how we can reduce tag ambiguity and spelling variations problems by taking into account social similarities calculated on folksonomies, in order to personalize resource recommendation.

In this context, a social personalized ranking function is proposed; this function merges the social aspect of folksonomy and events detection in order to estimate the relevance of given resources.

Keywords: Web 2.0, Semantic Web, folksonomies, Social Tagging, Social Interactions.

Résumé

Le tagging collaboratif qui est la clé des pratiques sociales du Web 2.0 a été fortement développé ces dernières années. La puissance des activités d'étiquetage social a permet aux utilisateurs d'ajouter des annotations sur les ressources en utilisant des tags. Ces tags expriment les intérêts des utilisateurs, leurs préférences et leurs besoins, mais aussi génèrent automatiquement des folksonomies. Les folksonomies ont été impliqués dans plusieurs approches de recherche d'informations et de recommandations. Cette thèse de doctorat aborde différents défis dans le domaine de web social afin d'exploiter l'aspect social des folksonomies dans les systèmes de recherche d'information et de recommandation.

Nous avons proposé une nouvelle approche pour analyser les profils des utilisateurs en fonction de leurs tags afin d'améliorer la recommandation des ressources. Nous nous sommes basés sur les règles d'association qui est une technique puissante de data mining pour découvrir des relations intéressantes entre les données du web. Notre objectif est de recommander des ressources annotées avec des étiquettes proposées par ces règles d'association afin d'enrichir les profils des utilisateurs. L'efficacité de la recommandation dépend de la résolution des problèmes d'étiquetage social. Dans notre processus de recommandation, nous montrons comment nous pouvons réduire l'ambiguïté des tags et le problème de variations orthographiques en tenant compte les similitudes sociales calculées sur les folksonomies afin de personnaliser la recommandation de ressources.

Dans ce contexte, une fonction de classement des ressources est proposée; cette fonction s'appuie sur l'aspect social et la détection des événements au sein des folksonomies pour estimer la pertinence des ressources proposées aux utilisateurs.

Mots clés : Web 2.0, Web Sémantique, Folksonomies, Tagging Social, Interactions sociales.

ملخص

الترميز التعاوني الذي هو حجر الأساس في الممارسات الاجتماعية للويب2.0 تم تطويره بشكل كبير في السنوات القليلة الماضية.

اقترحنا طريقة جديدة لتحليل ملفات تعريف المستخدمين وفقا لوضع علامات على النشاط من أجل تحسين إقتراح الموارد. لقد إرتكزنا على قواعد الارتباط التي تعد وسيلة قوية لاكتشاف علاقات مفيدة بين مجموعات كبيرة من البيانات على شبكة الإنترنت. هدفنا كان إقتراح الموارد بناء على السمات التي اقترحتها قواعد تكوين الإرتباط من أجل إثراء ملفات المستخدمين. فعالية هذا الإقتراح تعتمد على تجاوز عيوب العلامات الاجتماعية. علينا أن نظهر كيف يمكننا الحد من مشاكل علامة الغموض والاختلافات الإملائية من خلال مراعاة أوجه التشابه المحسوبة على الترميز التعاوني للمستخدمين ، و هذا من أجل تخصيص وتحسين إقتراح الموارد. لقد تغلبنا أيضا على عدم وجود وصلات دلالية بين علامات الترميز أثناء عملية الإقتراح. في هذا السياق، نقترح دالة ترتيبية للموارد المقترحة تستند على الترميز اليومي للمستخدمين.

كلمات البحث: ويب 2.0، ويب دلالي، الترميز التعاوني ،الترميز، التفاعلات الاجتماعية.

Publications

This thesis is based on the following original articles:

1. S. Beldjoudi, H. Seridi and C. Faron-Zucker. Personalizing and Improving Resource Recommendation by Analyzing Users Preferences in Social Tagging Activities, Accepted paper in Computiong and Informatics Journal, 2014 (To appear).
2. S. Beldjoudi, H. Seridi and C. Faron-Zucker. Personalizing and Improving Tag-based Search in Folksonomies. In Proc. Of the 15th International Conference on Artificial Intelligence Methodology, Systems, Applications (AIMSA), Springer LNAI 7557, pp. 112–118, 2012.
3. S. Beldjoudi, H. Seridi and C. Faron-Zucker. The Social Semantic web between the meaning and the mining. In Proc. Of the COSI 2012 conference (Colloque sur l'Optimisation et les Systèmes d'Information), 2012.
4. S. Beldjoudi, H. Seridi and C. Faron-Zucker. Let Tagging be more Interesting. In Proc. Of the IEEE Second International Workshop on Advanced Information Systems for Enterprises (IWAISE), 2012.
5. S. Beldjoudi, H. Seridi and C. Faron-Zucker. Ambiguity in Tagging and the Community Effect in Researching Relevant Resources in Folksonomies. In Proc. of ESWC workshop User Profile Data on the Social Semantic Web, 2011.
6. S. Beldjoudi, H. Seridi and C. Faron-Zucker. Improving Tag-based Resource Recommendation with Association Rules on Folksonomies. In Proc. of ISWC workshop on Semantic Personalized Information Management: Retrieval and Recommendation, 2011
7. S. Beldjoudi, H. Seridi and C. Faron-Zucker.les folksonomies entre la sémantique et l'effet communautaire. In Proc. Of JED, 2010.

Contents

Part II : Contributions

List of Figures

List of Tables

Chapter 1: Introduction

Chapter 1:
Introduction

The powerful of social tagging activities allows the wide set of web users to add free annotations on resources. Tags express user interests, preferences and needs, but also automatically generate folksonomies. These later can be considered as gold mine to provide effective information. Folksonomies have been also involved in many web searching and recommendations approaches, therefore it seems logical to use information about and derived from social network structures in the context of information retrieval systems. The presented PhD thesis addresses different challenges in the social web area. The main focus of our thesis is how to exploit social aspect of folksonomies within information retrieval, especially in searching and recommendation. This chapter explains the research context, motivations and questions we addressed in our study. Contributions are based on different approaches to solving the addressed research questions. In the last section, we present the organization of the presented thesis.

1.1 General Context

The social web has attracted the attention of millions of users as well as billions of Euros in these last years. As more social websites are formed around the connections between people and their objects of interest; and as these object-centered networks grow bigger and more diverse, more intuitive approaches are required for representing and navigating content within social platforms. There are two information access paradigms that users undertake each time they require to meet particular information needs on the web: searching by query and recommendation.

Querying a search engine is an effective approach that directly retrieves documents from an index of millions of documents in a fraction of a second. The other dominant information access paradigm involves recommendation-based systems that suggest items, such as movies, music or products by analyzing what the users with similar tastes have chosen in the past.

The simplicity of tagging combined with the culture of exchange allows the mass of users to share their annotations on the mass of resources. However, the exploitation of folksonomies raises several issues highlighted. This thesis will give a detailed description of the social information retrieval in the web, where new challenges are being tackled to overcome these limitations in a variety of social web application areas.

1.2 Thesis Context

In this thesis, substantial contributions related to social web technologies will be presented. Precisely, significant experiences in a number of applications in which folksonomies could be put to use will be discussed in detail. The presented contributions are based on information retrieval, results ranking and recommendation.

Our general purpose in this thesis was to capture, represent, and search from social web technologies to provide relevant personalized results to users. The main objectives are:

- Present a deep description of the state of the art in social web applications including: information retrieval, ranking and recommender systems in social web.
- Overcoming tag ambiguity and spelling variation problems in folksonomies.
- Improving resource recommendation within folksonomies.
- Propose a personalized social ranking function to rank retrieved resource within collaborative applications.

1.3 Motivation

With web 2.0 technologies, the web has become a social space where users create, annotate, share and make public resources which they find interesting on the web. Folksonomies are one of the keystones of these new social practices: they are systems of classification resulting from collaboratively creating and managing tags to annotate and categorize contents.

Despite the strength of folksonomies, there are some problems hindering the growth of these systems: tag ambiguity (or polysemy) is one of the famous problems in folksonomies. It comes from the fact that a tag can designate several concepts (i.e., a tag can have several meanings), for example when a user employs the tag "apple" to annotate a resource, the system will not understand if the user means the fruit or the company. Also the variations in writing a same concept (spelling variations or synonymy) can cause some problems during the search phase, for example "cat" and "chat" both denote the same concept (animal) in English and in French, but when a user searches resources annotated by the tag "cat", the system will not offer him those tagged with the word "chat" because it cannot understand that the tag "cat" has the same meaning that the tag "chat". In addition, tags that are freely chosen in these systems are likely to contain spelling errors and therefore make the retrieval of resources more uncertain than the metadata recovering from a word list examined by information professionals. Consequently resource retrieval within folksonomies needs some improvements to increase the quality of the results obtained in these systems.

The interest in recommender systems still remains high because it constitutes a problem-rich research area. Applying such systems within folksonomies recognized a real success in the last years, however due to the above-mentioned folksonomies problems it is clear that resource retrieval and so resource recommendation within folksonomies needs some improvements to increase the quality of results obtained in these systems. Consequently, it is interesting to analyze user profiles to overcome folksonomies problems and thus to improve the effectiveness of recommending personalized resources.

In folksonomies a relatively large number of resources can match users' queries. Therefore, ranking these web resources is a key problem, since a user cannot browse all them. Consequently, in this thesis we are interested by exploiting the power of social interactions between folksonomy users and event detection to improve resources retrieval in collaborative applications.

1.4 Research Questions and Issues

Based on the motivation stated in the previous section, our main research question is: how to improve recommendation and the retrieval effectiveness when searching personalized resources within social web applications?

We want to resolve the main research question by (1) analyzing users' profiles in social networks to help understanding tags' semantic during resource retrieval and recommendation, and (2) explicitly modeling the event dimension into retrieval and resources ranking.

Hence, the research questions we address are corresponding to two topics in social web: resources recommendation and ranking models within folksonomies. More specific research questions are presented below:

1.4.1 Resources recommendation in folksonomies

As it is motioned above, despite the strength of folksonomies, there are some problems hindering the growth of these systems: tag ambiguity, spelling variations and the lack of semantic links between tags.

Recommender systems have become an important research area since the mid-1990s. The interest in this domain still remains high because it constitutes a problem-rich research area.

Applying recommender systems within folksonomies recognized a real success in the last years, however due to the cited folksonomies problems it is clear that resource retrieval and so resource recommendation within folksonomies needs some improvements to increase the quality of the results obtained in these systems. The first research question we address is:

RQ1. How to analyze user profiles within folksonomies to overcome tag ambiguity and spelling variation problems and thus improve the effectiveness of recommending personalized resources?

Diabetes affects millions of people in the world leading to substantial negative effects and expensive healthy penalties in our life. Recently with the emergence of social networks in the internet and their use in different field, we want to use this technology in clinical practice by showing a system based on giving doctors relevant medical resources can be annotated by them. Thus the second research question we address is:

RQ2. How to help doctors discovering the best practices to patient diseases, diagnosis and treatments by analyzing doctors' profiles according to their daily tagging activity in order to personalize recommendation of medical resources?

1.4.2 Retrieval and Ranking Models

In many cases, when searching web resources, search results are displayed in chronological order where recently created resources are ranked higher than older ones. However, chronological ordering is not always effective. Therefore, a retrieval model should rank resources by the degree of relevance with respect to time. More precisely, documents must be ranked according to both social and temporal similarity.

An event-aware ranking model should take into account event detection, which captures the fact that the relevance of resources may change over time according to new events detection. Thus, the third research question we address is:

RQ3. How to explicitly model the event dimension into resources ranking?

In general, an event-aware ranking model gives scores to resources with respect to temporal feature. However, we want to study whether exploiting other features together with event detection can help improving the retrieval effectiveness in searching or recommending resources within social applications. In this case, we need to find features used for capturing the similarity between an information need and personalization of retrieved resources, and combine such features with event dimension for relevance ranking. Thus, the last research question we address in this thesis is:

RQ4. How to combine other features like social similarities with event detection in order to improve relevance ranking?

1.5 Main Contributions

This thesis is set in the research effort to improve social information retrieval on the web. In particular, we addressed social search and recommendation within web 2.0. Social tagging which is the keystone of the new social practices of web 2.0 has been highly developed in the last few years.

Firstly, we proposed a new method to analyzing user profiles according to their tagging activity in order to improve resource recommendation in folksonomies. We based upon association rules which are a powerful method to discover interesting relationships among large datasets on the web. Focusing on association rules we can find correlations between tags in a social network. Our aim is to recommend resources annotated with tags suggested by association rules, in order to enrich user profiles. The effectiveness of the recommendation depends on the resolution of social tagging drawbacks. In our recommender process, we demonstrate how we can reduce tag ambiguity and spelling variations problems by taking into account social similarities calculated on folksonomies in order to personalize resource recommendation. We surmount also the lack of semantic links between tags during the recommendation process. Moreover, a social personalized ranking function is proposed; this function leverages the social aspect of folksonomy and events detection to estimate the relevance of given resources to users.

In this context, a novel application is proposed to help doctors discovering the best practices to patient diseases, diagnosis and treatments in their daily tasks by analyzing doctors' profiles according to their tagging activity. We propose to take profit of community effect strength which characterizes social networks to show through an empirical scenario how we can evaluate and demonstrate the efficiency of social recommender system in clinical decision.

1.6 Thesis Organization

The remaining of this thesis is organized in five chapters:

Chapter 2 introduces the background needed as well as the basic concepts used throughout this thesis, including information retrieval, social web and social information retrieval which is defined as the bridge that fills the gap between information retrieval and the Web 2.0.

Chapter 3 presents a deep description of the state of the art in social information retrieval including: searching and recommendation. We focused on the most important concepts related to these two domains.

Chapter 4 presents our contribution to improve resource recommendation in folksonomies. We suggest an approach that considers social similarities calculated on folksonomies in order to personalize resource recommendation based on association rules. Also a social personalized ranking function is proposed to leverage the social aspect of folksonomy and events detection. The aim is to acquire relevant resources recommended to or inquired by users.

Chapter 5 describes details of implementing the proposed approaches in order to evaluate and demonstrate their efficiency. Also, we implemented a world application for diabetes disease, which put into practice these approaches.

Chapter 6 presents conclusions and research perspectives.

Part 1:
State of the Art

Chapter 2: Research Issues and Background

Chapter 2

Research Issues and Background

This chapter reviews the main concepts addressed in this thesis. Firstly, we present the background knowledge required to understand information retrieval and its basic process. In section 2, the social web and the main models of social relationships will be introduced; this includes specially folksonomies in the context of web 2.0. In the last section, we introduce social information retrieval, which is defined as the bridge that fills the gap between information retrieval and the Web 2.0.

2.1 Information retrieval

Information Retrieval (IR) is the art of presentation, organization of and access to information items. The representation and arrangement of information should be in such a manner that the user can access information to find his information need. The definition of IR according to [Manning et al., 2009] is:

Information retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

In the following subsections, we will introduce the basic notions related to this domain in more details.

2.1.1 Difference between information retrieval and data retrieval

As we are moving into the information age there is a requirement to protect data transmission and also to store and retrieve the data and the information. Table 2.1 gives an overview about the basic differences between these two concepts.

DATA	INFORMATION
It is the raw fact.	Information is processed data
For its retrieval it needs to be fully mentioned. If the file name or the document name is not known or is case sensitive, there are chances for the system to fail and do not retrieve any document.	For its retrieval a partial information is enough for its evaluation.
Examples of data are *a piece of paper, a book, an algorithm*.	Examples of information are *a piece of paper on a table, a book in the shelf, a*

	bubble-sort algorithm.
In the above examples their location is not known and hence the meaning cannot be given to this data.	In the above examples their location are known and hence they have a specified meaning.

Table 2.1: The main difference between data and information

Data retrieving systems can be found in the operating system search. Windows search is a best example for the data retrieval system. We would have to specify the exact name of the file that we desire. Where information retrieval systems are like web search engines. The best well-known is Google. It treats the natural language and generates the result including the most set of documents matching the query.

Recently, it became very important to retrieve the data in a faster manner. Previous there used to linear search mechanisms where the complete set of documents in the database are read and then sorted according to the query and displayed. This had varied complexities and took greater time compared to the advanced techniques available these days.

In an information retrieval system, the documents are scanned for the query. To decrease the computation time of the system, the documents are scanned only for the repeated key words which are considered relevant to the document. The result displayed sends a feedback as an input for the next query. In this way with every query there is an increase in the performance of the system.

The difference between an information retrieval system and a data retrieval system is that:

– IR deals with unstructured/semi-structured data while data retrieval (a database management system or DBMS) deals with structured data with well-defined semantics.

– Querying a DBMS system produces exact/precise results or no results if no exact match is found.

– Querying an IR system produces multiple results with ranking. Partial match is allowed document characterization.

In the next section, a description of IR Systems modules will be given.

2.1.2 Components of an Information Retrieval System

In this section we describe the components of a basic web information retrieval system. A general information retrieval functions in the following steps. It is shown in Figure 2.1.

- The system browses the document collection and gets documents. (Crawling step)
- The system builds an index of these documents. (Indexing step)
- User specifies his query.

- The system retrieves documents that are relevant to the query from the index and displays that to the user. (Ranking step)

- User can give relevance feedback to the search engine. (Relevance Feedback step)

The goal of any information retrieval system is to satisfy user's information need. Unfortunately, characterization of user information need is not simple. User's often do not know clearly about the information need. Query is only a vague and incomplete description of the information need. Query operations like query expansion, stop word removal etc. are usually done on the query.

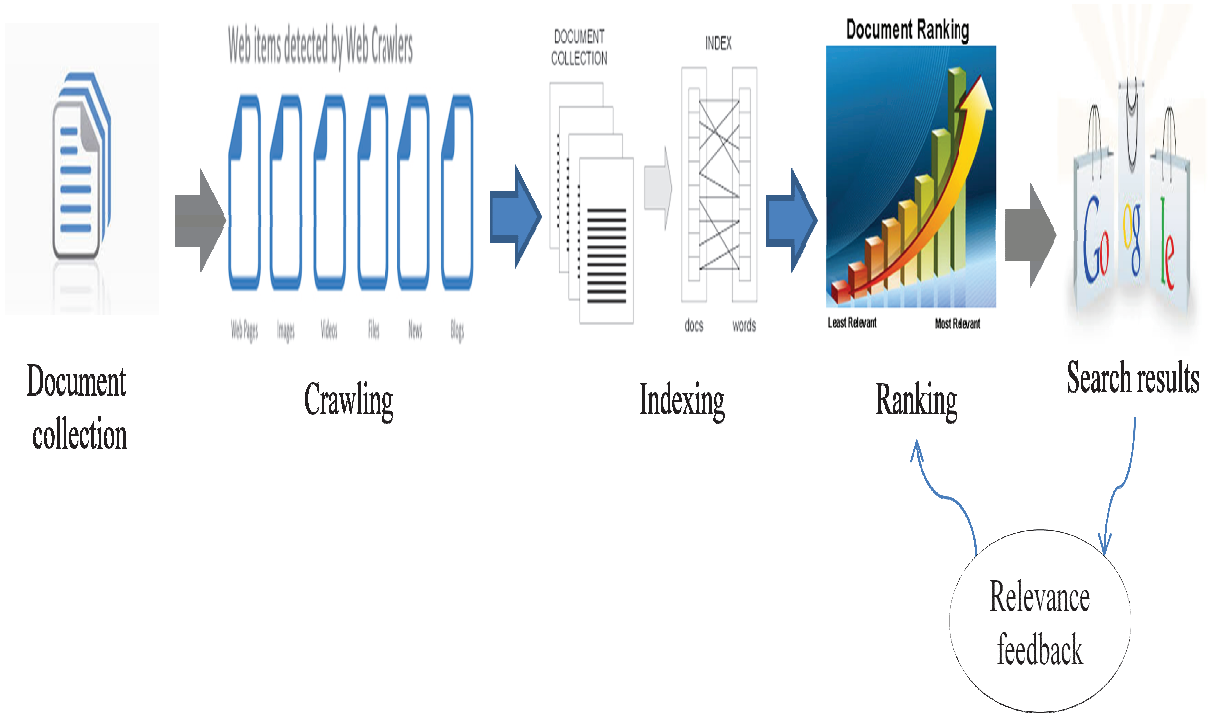


Figure 2.1: Important Processes in web information retrieval

Crawling

The web crawler retrieves documents from the web by following some defined strategies. The crawler creates a copy of all the documents it crawls to be treated by the search engine. The crawler starts from a list of documents called seed. The crawler visits the documents (URLs), identifies the outgoing hyperlinks there and adds them to the list of documents to be visited. This way the crawler traverses the web graph following hyperlinks. It saves a copy of each document it visits.

a) Selection policy

Selection policy decides the links to crawl first. Usually the web graph is traversed in a breadth first way to avoid being lost at infinite depth. As the number of documents is enormous, the selection policy becomes important so as to select which documents to crawl and which documents not to crawl. Generally page importance measures like PageRank are used as a selection strategy.

b) Revisit policy

The crawler needs to crawl frequently to keep the search results up-to-date. The revisit policy determines how frequently the crawling process should be restarted. There is a cost associated with an outdated copy of a document. The mostly used cost functions are freshness (is the stored copy outdated?) and age (how old is the stored copy). There may be two revisit policies:

Uniform policy revisit all the documents in the collection with same frequency and proportional policy revisit documents that change frequently more often.

It is interesting to note that proportional policy often incurs more freshness cost. The reason being, pages in the web either keep static or change so frequently that even the proportional policy cannot keep them up to date.

Indexing

The documents crawled by the search engine are stored in an index for efficient retrieval. The documents are first parsed, and then tokenized, stop-word removed and stemmed. After that they are stored in an inverted index. The process is discussed below.

a) Tokenization

This stems extracts word tokens (index terms) from running text. For example, given a piece of text: “Places to be visited in Algeria” it outputs [places, to, be, visited, in, Algeria].

b) Stop-word eliminator

Stop-words are those words that do not have any disambiguation power. Common examples of stop words are articles, prepositions etc. In this step, stop words are removed from the list of tokens. For example, given the list of token generated by tokenizer, it strips it down to: [places, visited, Algeria].

c) Stemmer

The remaining tokens are then stemmed to the root form (e.g. visited → visit). For example, after stemming the list of tokens becomes this: [place, visit, Algeria].

d) Inverted index

The ordinary index would contain for each document, the index terms within it. But the inverted index stores for each term the list of documents where they appear. The benefit of using an inverted index comes from the fact that in IR we are interested in finding the documents that contain the index terms in the query. So, if we have an inverted index, we do not have to scan through all the documents in collection in search of the term. Often a hash-table is associated with the inverted index so that searching happens in O (1) time.

Inverted index may contain additional information like how many times the term appears in the document, the offset of the term within the document? etc.

Ranking

When the user gives a query, the index is consulted to obtain the documents most relevant to the query. The relevant documents are then ranked according to their degree of relevance, importance, etc.

a) Relevance Feedback

Relevance feedback is one of the classical ways of refining search engine rankings. It works in the following way: Search engine firsts generate an initial set of rankings. Users select the relevant documents within this ranking. Based on the information in these documents a more appropriate ranking is presented (for example, the query may be expanded using the terms contained in the first set of relevant documents).

Sometimes users do not enough domain knowledge to form good queries. But they can select relevant documents from a list of documents once the documents are shown to him. For example, when the user fires a query 'matrix', initially documents on both the topics (movie and maths) are retrieved. Then say, the user selects the maths documents as relevant. This feedback can be used to refine the search and retrieve more documents from mathematics domain.

- *Types of relevance feedback*

Explicit: User gives feedback to help system to improve.

Implicit: User doesn't know he is helping e.g. "similar pages" features in Google.

Pseudo: User doesn't do anything! Top 'k' judgments are taken as relevant. Being fully automated it has always this risk that results may drift completely away from the intended document set.

- *Issues with relevance feedback*

The user must have sufficient knowledge to form the initial query.

This does not work too well in cases like: Misspellings and Mismatch in user's and document's vocabulary.

Relevant documents has to be similar to each other (they need to cluster) while similarity between relevant and non-relevant document should be small. That is why this technique does not work too well for generic topics that often appear as disjunction of more specific concepts.

Long queries generated may cause long response time.

Users are often reluctant to participate in explicit feedback. [Spink et al., 2000: Only 4% of users participate. 70% doesn't go beyond first page.]

2.1.3 Theoretical Models in Information Retrieval

Many retrieval models have been proposed, for example, Boolean retrieval model, vector space model, probabilistic model, language modeling approaches and learning-to-rank. Retrieval models differ from each other in several aspects like query interpretation, document representation, etc. In the following, we will give details to each of the retrieval models.

Boolean Retrieval Model

The Boolean retrieval model is the simplest IR model. A query is a mixture of terms and Boolean operators AND, OR and NOT. A document is modeled as bag of words where each term is represented using binary weighting {1, 0} (1 for term presence and 0 for term absence).

The Boolean retrieval model ignores the degree of relevance since it assumes two outcomes of relevance, i.e., relevant or non-relevant. Intuitively, the model returns all documents exactly matched with the query terms without ordering the documents.

Despite its simplicity, the retrieval effectiveness of a Boolean query depends entirely on the user. In order to obtain high effectiveness, the user can issue a complex query, but it is quite difficult to formulate. If a simple query is used, there might be too few or too many documents retrieved. If a large number of documents are retrieved, this poses a problem for the user because he has to spend time looking for those satisfying the information needs.

Vector Space Model

The vector space model is a ranked retrieval model. That is, documents are retrieved and ranked descendingly by the degree of relevance, which can be measured as the similarity between a query and a document. First, a query and documents are represented as vectors of term weights by using a term weighting scheme, e.g., *tf-idf*. Given a term *w* and a document *d*, *tf* is the term frequency of *w*, which is normalized by the total term frequency in *d*. Thus, *tf* can be computed as:

$$tf(w,d) = \frac{freq(w,d)}{\sum_{j=1}^{n_d} freq(w_j,d)} \qquad (2.1)$$

where *freq(w, d)* is the term frequency of *w* in *d* and n_d is the number of distinct terms in *d*. *tf* captures the importance of a term *w* in a document by assuming that the higher *tf* score of *w*, the more importance of *w* with respect to *d*. Intuitively, terms that convey the topics of a document should have high values of *tf*.

idf is the inverse document frequency weight of a term *w*. It measures the importance of *w* with respect to a document collection. *idf* can be seen as a discriminating property, where a term that appears in many documents is *less discriminative* than a term appears in a few documents. *idf* can be computed as:

$$idf(w) = \log \frac{N}{n_w} \qquad (2.2)$$

Where *N* is the total number of documents in a collection, and n_w is the number of documents in which a term *w* occurs. Finally, a *tf-idf* weight of a term *w* in a document *d* can be computed using the function *tf-idf (w,d)* given as:

$$tf\_idf(w,d) = tf(w,d).idf(w) \qquad (2.3)$$

Finally, a query *q* and a document *d* can be represented as vectors of *tf-idf* weights of all terms in the vocabulary as:

$$\vec{q} = \langle \psi_{1,q}, \dots, \psi_{n,q} \rangle \qquad (2.4)$$

$$\vec{d} = \langle \psi_{1,d}, \dots, \psi_{n,d} \rangle \qquad (2.5)$$

Where $\psi_{i,q}$ is *tf-idf* weight of a term w_i in *q* and $\psi_{i,d}$ is *tf-idf* weight of a term w_i in *d*.

The similarity of the term-weight vectors of *q* and *d* can be computed using the cosine similarity as:

$$sim(\vec{q},\vec{d}) = \frac{\vec{q}.\vec{d}}{|\vec{q}| \times |\vec{d}|} = \frac{\sum_{i=1}^{n} \psi_{i,q} \times \psi_{i,d}}{\sqrt{\sum_{i=1}^{n} \Psi_{i,q}^2 \times \sum_{i=1}^{n} \Psi_{i,d}^2}} \qquad (2.6)$$

The advantages of the vector space model over the Boolean retrieval model are: 1) it employs term weighting which improves the retrieval effectiveness, 2) the degree of similarity allows partially matching documents to be retrieved, and 3) it is fast and easy for implementing. However, there are some disadvantages of the vector space model. First, it makes no assumption about term dependency, which might lead to poor results [Baeza-Yates and Ribeiro-Neto, 2011]. In addition, the vector space model makes no explicit definition of relevance. In other words, there is no assumption about whether relevance is binary or mutivalued, which can impact the effectiveness of ranking models.

Probabilistic Model

The probabilistic model was first proposed by Robertson and Jones [Robertson and Jones, 1976]. The model exploits probabilistic theory to capture the uncertainty in the IR process. That is, documents are ranked according to the probability of relevance. There are two assumptions in this model: 1) relevance is a binary property, that is, a document is either relevant or non-relevant, and 2) the relevance of a document does not depend on other documents.

Given a query q, let R and $\bar{R}$ be the set of relevant documents and the set of non-relevant documents with respect to q respectively. A basic task is to gather all possible evidences in order to describe the properties of the sets of relevant documents and non-relevant documents. The similarity of q and a document d can be computed using the odd ratio of relevance as:

$$sim(d, q) = \frac{P(R|d)}{P(\bar{R}|d)} \qquad (2.7)$$

In order to simplify the calculation, Bayes' theorem is applied yielding the following formula:

$$sim(d, q) = \frac{P(R|d)}{P(\bar{R}|d)} = \frac{P(R).P(d|R)}{P(\bar{R}).P(d|\bar{R})} \approx \frac{P(d|R)}{P(d|\bar{R})} \qquad (2.8)$$

Where $P(R)$ is the prior probability of a relevant document, and $P(\bar{R})$is the prior probability of a non-relevant document. For a given query q, it is assumed that both prior probabilities are the same for all documents, so they can be ignored from the calculation.

$P(d|R)$ and $P(d|\bar{R})$ are probabilities of randomly selecting a document d from the set of relevant documents R and the set of non-relevant documents $\bar{R}$ respectively.

In the probabilistic model, a document d is represented as a vector of terms with binary weighting, which indicates term occurrence or non-occurrence.

$$\vec{d} = \langle \psi_{1,d}, \dots, \psi_{n,d} \rangle \qquad (2.9)$$

Where $\psi_{i,d}$ is the weight of a term w_i in a document d, and $\psi_{i,d} \in \{0, 1\}$. In order to compute $P(d|R)$ and $P(d|\bar{R})$, it assumes the Naive Bayes conditional independence [Manning et al., 2009], that is, the presence or absence of a term in a document is independent of the presence or absence of other terms in the given query. Thus, the computation of similarity can be simplified as:

$$sim\ (d, q) \approx \frac{P(d|R)}{P(d|\bar{R})} \approx \frac{\prod_{i=1}^{n} P(w_i|R)}{\prod_{i=1}^{n} P(w_i|\bar{R})} \qquad (2.10)$$

Where $P(w_i|R)$is the probability that a term w_i occurs in relevant documents, and $P(w_i|\bar{R})$ is the probability that a term w_i occurs in non-relevant documents. By modeling relevance using probability theory makes the probabilistic model theoretically sound compared to the Boolean retrieval model and the vector space model. However, a drawback is an independence assumption of terms, which is contrary to the fact that any two terms can be semantically related. In addition, the probabilistic model is difficult to implement because the complete sets of relevant documents and non-relevant documents are not easy to obtain. Thus, in order to compute $P(w_i|R)$ and $P(w_i|\bar{R})$, it is needed to guess prior probabilities of a term w_i by retrieving top-n relevant documents and then perform iterative retrieval in order to recalculate probabilities. This makes it difficult to implement the model. In addition, the probabilistic model ignores the frequency of terms in a document.

Language Modeling

Originally, language modeling was employed in speech recognition for recognizing or generating a sequence of terms. In recent years, language model approaches have gained interests from the IR community and been applied for IR. A language model M_D is estimated from a set of documents D, which is viewed as the probability distribution for generating a sequence of terms in a language. The probability of generating a sequence of terms can be computed by multiplying the probability of generating each term in the sequence (called a unigram language model), which can be computed as:

$$P(w_1, w_2, w_3|M_D) = P(w_1|M_D).P(w_2|M_D).P(w_3|M_D) \qquad (2.11)$$

The original language modeling approach to IR is called the *query likelihood model* [Manning et al., 2009].

In this model, a document d is ranked by the probability of a document d as the likelihood that it is relevant to a query q, or $P(d|q)$. By applying Bayes' theorem, $P(d|q)$ can be computed as:

$$P(d|q) = \frac{P(q|d).P(d)}{P(q)} \qquad (2.12)$$

Where $P(q)$ is the probability of a query q, and $P(d)$ is a document's prior probability.

Both $P(q)$ and $P(d)$ are in general ignored from the calculation because they have the same values for all documents. The core of the *query likelihood model* is to compute $P(q|d)$ or the probability of generating q given the language model of d, M_D. $P(q|d)$ can be computed using maximum likelihood estimation (MLE) and the unigram assumption where n_q is the number of terms in q. The equation above is prone to zero-probability, which means that one or more terms in q may be absent from a document d. In order to avoid zero-probability, a smoothing technique can be applied in order to add a small (non-zero) probability to terms that are absent from a document. Such a small probability is generally taken from the background document collection. For each query term *w*, a smoothing technique is applied yielding the estimated probability $\hat{p}(w|d)$ of generating each query term *w* from *d* as:

$$\hat{P}(w|d) = \lambda\,.\,P(w|M_d) + \ (1-\lambda).\,P(w|M_c) \qquad (2.13)$$

Where the smoothing parameter $\lambda \in [0, 1]$. C is the background document collection. M_c is the language model generated from the background collection.

Learning to Rank

Many researchers have applied machine learning algorithms in order to optimize the quality of ranking, called learning-to-rank approaches. In general, there are three main steps for modeling a ranking function using learning-to-rank approaches [Liu, 2009]:

1. *Identify features*. A set of features $\{x_1, x_2, . . . , x_m\}$ are defined as sources of the relevance of a document d_i with respect to a query q_j . Normally, a value of each feature x_i is a real number between [0, 1]. The same notation will be used for both feature and its value that is x_i. Given a query q_j , a document d_i can be represented as a vector of feature values, $d_i = (x_1, x_2, . . . , x_z)$ indicating the relevance of d_i with respect to q_j .

2. *Learn a ranking model*. Machine learning is used for learning a ranking function *h(q, d)* based on training data, called supervised learning. Training data is a set of triples of labeled or judged query/document pairs $\{(q_j\ ,\ d_i,\ y_k)\}$, where each document d_i is represented by its feature values, $i \in \{1, . . . , n\}$ and $j \in \{1, . . . ,m\}$. A judgment or label y_k can be either relevant or non-relevant $y_k \in \{1,-1\}$, or a *rank* representing by natural numbers $y_k \in$ N.

3. *Rank documents using models*. The ranking function *h(q, d)* learned in the previous step will be used for ranking test data, or a set of unseen query/document pairs $\{(q_j\ ,\ d_i)\}$ where $i \notin \{1, . . . , n\}$ and $j \notin \{1, . . . ,m\}$. The result is a judgment or label y'_k for each query/document pair.

A ranking model *h(d, q)* is obtained by training a set of labeled query/document pairs using a learning algorithm. A learned ranking model is essentially a weighted coefficient w_i of a feature x_i. An unseen document/query pair (d′, q′) will be ranked according to a weighted sum of feature scores:

$$score\ (d', q') = \sum_{i=1}^{N} w_i \times\ x_i^{q'} \qquad (2.14)$$

Where *N* is the number of features.

Many existing learning algorithms have been proposed, and can be categorized into three approaches: pointwise, pairwise, and listwise approaches [Liu, 2009]. The pointwise approach assumes that retrieved documents are independent, so it predicts a relevance judgment for each document and ignores the positions of documents in a ranked list. The pairwise approach considers a pair of documents, and relevance prediction is given as the relative order between them (i.e., pairwise preference). The listwise approach considers a whole set of retrieved documents, and predicts the relevance degrees among documents [Liu, 2009].

2.1.4 Temporal Information Retrieval

Temporal information retrieval refers to IR tasks that analyze and exploit the time dimension embedded in documents to provide alternative search features and user experience.

Basically, two types of temporal information particularly useful for temporal IR: 1) the publication or creation time of a document, and 2) temporal expressions mentioned in a document or a query. In the following, we first give an overview of different types of temporal expressions. Then, we present time models.

Temporal Expressions

As explained in [Alonso et al., 2007], there are three types of temporal expressions: explicit, implicit and relative. An explicit temporal expression mentioned in a document can be mapped directly to a time point or interval, such as, dates or years on the calendar. For example, “July 05, 1962” is an explicit temporal expression.

An implicit temporal expression is given in a document as an imprecise time point or interval. For example, “Independence Day 1962” is an implicit expression that can be mapped to “July 05, 1962”.

A relative temporal expression occurring in a document can be resolved to a time point or interval using a time reference - either an explicit or implicit temporal expressions mentioned in a document or the publication date of the document itself. For example, the expressions “this Monday” or “next month” are relative expressions which we map to exact dates using the publication date of the document.

Models for Time, Documents and Queries

In temporal IR, the time dimension must be explicitly modeled in documents and queries. In the following, we outline models for time, documents and queries that are employed in temporal IR tasks.

a) **Time Models**

[de Jong et al., 2005] modeled time as a *time partition*, that is, a document collection is partitioned into smaller time periods with respect to a time granularity of interests, e.g. *day*, *week*, *month*, or *year*. A document collection C contains a number of corpus documents, such as, $C = \{d_1, \ldots, d_n\}$. A document d_i is composed of bag-of-words, and the publication time of

d_i is represented as $Time(d_i)$. Thus, d_i can be represented as $d_i = \{\{w_1, . . . ,w_n\} , Time(d_i)\}$. Given a time granularity of interest and C is partitioned into smaller time periods, the associated time partition of d_i is a time period $[t_k, t_{k+1}]$ that contains the publication time of d_i, that is $Time(di) \in$ [tk, tk+1]. For example, if the time granularity of *year* is used, the associated time interval for 2015/03/08 will be [2015/01/01, 2015/12/31].

[Berberich et al., 2010] represented a temporal expression extracted from a document or the publication time of a document as a quadruple: (tb_l, tb_u, te_l, te_u) where tb_l and tb_u are the lower bound and upper bound for the begin boundary of the time interval respectively, which underline the time interval's earliest and latest possible begin time.

Similarly, te_l and te_u are the lower bound and upper bound for the end boundary of the time interval respectively, which underline the time interval's earliest and latest possible end time. Since the time interval is not necessarily known exactly, the time model of [Berberich et al., 2010] is proposed to capture lower and upper bounds for the interval boundaries.

To interpret the time uncertainty in this model, consider the following example given in [Berberich et al., 2010]. The temporal expression "in 2015" is represented as (2015/01/01, 2015/12/31, 2015/01/01, 2015/12/31), which can refer to any time interval $[b, e]$ having a begin point $b \in [tb_l, tb_u]$ and an end point $e \in [te_l, te_u]$ where $b \leq e$. Note that, the actual value of any time point, e.g., tb_l, tb_u, te_l, or te_u, is an integer or the number of time units (e.g., milliseconds or days) passed (or to pass) a reference point of time.

These time units are referred as *chronons* and a temporal expression t is denoted as the set of time intervals that t can refer to.

b) Document Model

A document d consists of a textual part d_{text} (an unordered list of terms) and a temporal part d_{time} composed of the publication date and a set of temporal expression $\{t_1, . . . t_k\}$. The publication date of d can be obtained from the function *PubTime(d)*. Temporal expressions mentioned in the contents of d can be obtained from the function *ContentTime(d)*. Both the publication date and temporal expressions can be represented using the time models defined above.

c) Temporal Query Model

A temporal query q refers to a query representing *temporal information needs*, which is composed of two parts: keywords q_{text} and a temporal expression q_{time}. In other words, a user wants to know about documents that are relevant to both the topic of interest and temporal intent. Temporal queries can be categorized into two types: 1) those with time explicitly specified, and 2) those with implicit temporal intents. An example of a query with time explicitly specified is the occupation of Algeria in 1830.

In this case, a temporal intent is represented by the temporal expression “before 1830” indicating that a user wants to know about colonization events in Algeria in the year 1830. Similarly, the temporal part of a query or q_{time} can be represented using any time models defined above.

Note that, there is no standard terminology for referring a query in this research area. Previous work [Kulkarni et al., 2011, Metzler et al., 2009, Nørvåg, 2004, Wang et al., 2010] mainly uses the term *temporal queries*, however, the term *time-sensitive queries* has been used recently in some work [Dakka et al., 2008, Dong et al., 2010, Zhang et al., 2009].

2.1.5 Evaluation of IR Systems

Here we discuss different evaluation measures of Information Retrieval techniques. We first describe standard measures like precision and recall. Then we discuss different combined measures.

Precision

Precision measures the exactness of the retrieval process. If the actual set of relevant document is denoted by *I* and the retrieved set of document is denoted by *O*, then the precision is given by:

$$precision = \frac{|I \cap O|}{|O|} \qquad (2.15)$$

Precision measures among the retrieved documents, how many are relevant. It does not care if we do not retrieve all the relevant documents but penalizes if we retrieve non-relevant documents.

Precision at Rank k

In web search systems the set of retrieved document is usually huge. But it makes a lot of difference if the relevant document is retrieved early in the rank list that late. To take this into account, precision at a cut-off rank *k* is introduced. Here the list of relevant documents *I* is cut-off at rank *k*. Only documents up to rank *k* are considered to be retrieved set of documents.

Mean average precision or MAP score

To represent the order in which the result was given, the mean average precision or *MAP* measure gives the average of precision at various cut-off ranks. It is given by:

$$MAP = \frac{\sum_{i=0}^{k} precision@i}{k} \qquad (2.16)$$

Recall

Recall is a measure of completeness of the IR process. If the actual set of relevant document is denoted by *I* and the retrieved set of document is denoted by *O*, then the recall is given by:

$$recall = \frac{|I \cap O|}{|I|} \qquad (2.17)$$

Recall measures how much of the relevant set of documents we can retrieve. It does not care if we retrieve non-relevant documents also in the process.

Precision and Recall are not independent measures. It is seen that if we try to increase precision, the recall is reduced and vice versa.

F-Score (F1 metric)

F-Score tries to combine the precision and Recall measure. It is the harmonic mean of the two. If *P* is the precision and *R* is the recall then the F-Score is given by:

$$F1 = \frac{2.P.R}{P + R} \qquad (2.18)$$

Information Retrieval Evaluation Forums

IR Evaluation forums provide framework to test various IR techniques against standard benchmarks. They provide the following things: [Manning et al., 2009]

- A Document collection
- A test suite of information needs expressible as queries
- A set of relevance judgments. Usually, a set of documents is manually judged as relevant or non-relevant.

One key thing to note is that relevance is judged against information needs and not queries [Manning et al., 2009]. The task of converting an Information Need to a system query must be done by the system itself and becomes a part to be evaluated [Baeza-Yates and Ribeiro-Neto, 1999]. Here we discuss four such evaluation forums namely TREC[1], CLEF[2], FIRE[3] and NTCIR[4].

[1] http://trec.nist.gov/

[2] http://www.clef-campaign.org/

[3] http://www.isical.ac.in/~fire/

[4] http://research.nii.ac.jp/ntcir/index-en.html

a) TREC

TREC stands for Text REtrieval Conference. It started in 1992. The purpose of TREC is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. Also increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas. In addition TREC aimed at speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems and increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

TREC is divided in different research areas called TREC Tracks. These tracks act as incubators of new areas and creates necessary infrastructure for the area. Each track has a task to be performed. Multiple groups across the globe participate to do the task. The participating groups are provided with data sets and test problems. Each Track has an open-to-all mailing list to discuss the task. Tracks may be added or removed from TREC depending on changing research needs.

TREC retrieval tasks are to be performed with reference to certain test information requests (needs) specified in TREC topics. Each topic is a description of information need in natural language. The task of convert a topic into a system query must be done by the system itself [Baeza-Yates and Ribeiro-Neto, 1999].

b) CLEF

CLEF or Cross Language Evaluation Forum is an IR evaluation forum dealing in multilingual information retrieval in European languages. It was founded by European Union in 2000 [Manning et al., 2009]. The aims of CLEF according to its official website are developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts and also creating test-suites of reusable data which can be employed by system developers for benchmarking purposes.

The evaluations are concentrated on cross-language information retrieval in European languages.

Like TREC, CLEF is also divided into topics (tracks) that researches on different areas. The key areas are: searching on a text (ad-hoc task), geographical information search (GeoCLEF), search of information on the web (WebCLEF), image retrieval (ImageCLEF), question answering systems (QA@CLEF), etc. Each track offers a task (problem) which is relevant to that research area. A collection of objects (documents, news, images etc.) is given to the participants. They are also given certain search objectives (topics) and queries. Based on these, the participants build indexes and the ranking system. They execute their searches and send the result in a standard format. The organizers select the best groups based on the precision they achieve.

c) FIRE

Forum for Information Retrieval Evaluation (FIRE) is an evaluation forum coordinated by Indian Statistical Institute, Kolkata. The goals of FIRE can be summarized in encouraging research in South Asian language Information Access technologies by providing reusable large-scale test collections for experiments. Also exploring new Information Retrieval / Access tasks that arise as our information needs evolve, and new needs emerge. Provide a common evaluation infrastructure for comparing the performance of different IR systems

d) NTCIR

NII Test Collection for IR Systems (NTCIR) is coordinated by National Institute of Informatics (NII), Japan. It is an information retrieval evaluation forum like TREC or CLEF, specifically focused towards East Asian languages like Japanese, Chinese and Korean. The goals as given in the official website are:

- To encourage research in Information Access technologies by providing large-scale test collections reusable for experiments and a common evaluation infrastructure allowing cross-system comparisons
- To provide a forum for research groups interested in cross-system comparison and exchanging research ideas in an informal atmosphere
- To investigate evaluation methods of Information Access techniques and methods for constructing a large-scale data set reusable for experiments.

As this thesis is concerned with the improvement of social information retrieval, in the next section we will introduce some related concept to social web.

2.2 Social Web and Folksonomies

In this part, a general overview about social web technologies, especially folksonomies will be presented.

2.2.1 Web 2.0 definition

Web 2.0 describes the second generation of web sites that use more technologies further than the static pages of the previous web sites. This term was invented in 1999 by Darcy DiNucci and was popularized by Tim O'Reilly in 2004.

In software engineering, new versions of programs are named with an incremental version number. The notation "2.0" comes from this principle; this novel generation of the web takes account of new features that did not exist in the past. However, web 2.0 is not a new version of the web; it is rather a series of technological improvements.

Web 2.0 technologies allow users to interact and create their own contents in a virtual community, in contrast to web 1.0 sites (Figure 2.2) where users are only limited to the passive viewing of content.

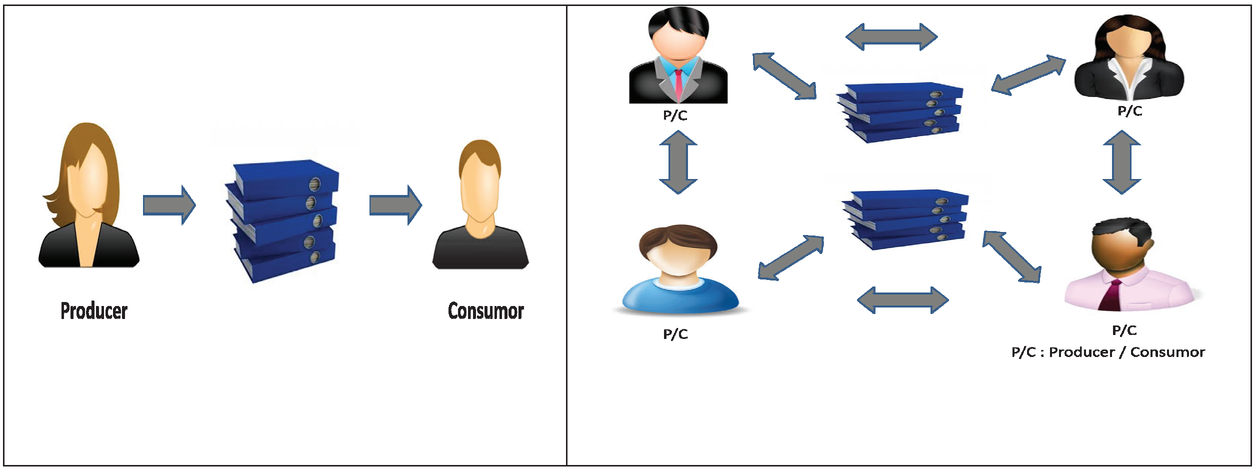


Figure 2.2: Web 1.0 vs. Web 2.0

2.2.2 Web 2.0 services and applications

There are a number of Web-based services and applications. Most of these applications are mature, having been in exploit for some years, although new features are being added on a regular basis. In the following section we introduce these popular used services.

Blogs

The concept web-log, or blog, was invented by Jorn Barger in 1997 and refers to a simple webpage consisting of short paragraphs of information, personal view, comments or links, called posts, arranged chronologically with the most recent first [Doctorow et al., 2002].

A blog is usually maintained by a single user or a small set of contributors. Visitors to the blog can comment or respond to comments made by other visitors. Blogs can also include photos, images, sounds, or videos.

Wikis

A wiki is a single or set of web pages that can be edited by anyone who is allowed access [Ebersbach et al., 2006]. The popular success of Wikipedia has meant that the concept of the wiki, as a collaborative tool that assists the production of a group work, is extensively understood. Wiki pages have an edit button allows users editing, change or even delete the contents of the page in question. Simple, hypertext links between pages are used to generate a navigable set of pages. Wikis can also include images, sound and videos.

Different to blogs, wikis usually have a history function, that allows earlier versions to be examined, and a rollback function, that restores previous versions.

Tagging and social bookmarking

A tag is a keyword that is added to a web resource (e.g. a website, photo, video, etc.) to describe it. Tagging concept has been extended far beyond website bookmarking, and services like Flickr, YouTube, and Odeo allow a variety of digital web resources to be socially tagged. The idea of tagging has been expanded to include what are called tag clouds that are groups of tags from a number of different users of tagging service, which collates information about the tags frequency. This information is often displayed graphically as a cloud in which tags with higher frequency of use are displayed in bigger text.

Multimedia sharing

One of the principal increase areas has been amongst services that facilitate the storage and sharing of multimedia content. These services take the idea of writeable web where users are not just consumers but contribute actively to the production of Web content and enable it on a massive scale. Actually a big number of users contribute in the sharing and exchange of these forms of media by producing their own podcasts, videos and photos. This progress has only been made possible through the extensive adoption of high quality, but relatively low cost digital media technology such as hand-held video cameras.

Podcasting

Podcasting is a method to let audio or video files accessible in the internet that can either be exploited on a personal computer or downloaded to a hand-held tool such as an iPod or mp3 player. A podcast will be treated as an audio podcasts or a video podcasts. It may also comprise photos, like PowerPoint presentations. There may be several levels of copyright in a podcast, depending on the content of this later. If there is a broadcaster or a subject being interviewed, they will not only own copyright in their presentation or interview but they will have performers’ rights. There will also be a separate copyright in the actual recording itself.

Each of these services will have copyright implications that require to be managed. The copyright issues will differ depending on the kind of content that can be produced or contributed and how users of the site will act with the material on the site.

RSS and syndication

Certainly, in its earliest picture, RSS was understood to stand for Rich Site Summary [Doctorow, 2002]. For a multiplicity of historical causes, there are a number of RSS formats (RSS 0.91, RSS 0.92, RSS 1.0, RSS 2.0) and there are some issues of incompatibility. It is worth noting that RSS 2.0 is not simply a later version of RSS 1.0, but is a different format.

This is a web based system or solution that allows information from a web site to be subscribed to by other persons. It allows the data from that web site to be automatically 'pumped' out to subscribers of that web site. The data is referred to as a feed or a channel.

RSS feeds can be obtained from many different online sources and web sites including: Blogs, wikis, podcast, etc.

Subscribing to sites via RSS allows one to focus on the topics and subjects that they find interesting and purposeful.

RSS is a set of formats which let users find out about updates to the content of RSS-enabled websites, blogs or podcasts without in fact having to go and visit the site. Instead, information from the website is composed within a feed (which uses the RSS format) and piped to the user in a procedure recognized as syndication.

Precisely, RSS is an XML-based data format for websites to swap files that contain publishing information and summaries of the site's contents.

In this thesis, we are interested by a famous technology of social bookmarking which is folksonomies. In the following section a detailed description of this technology will be given.

2.2.3 Folksonomies

In the following subsections, a general overview about folksonomies will be presented:

Origin and definition

A folksonomy is a classification system derived from the process of collaboratively creating and exploiting tags to annotate web resources; this practice is also recognized as collaborative and social tagging.

This term introduced by Thomas Vander Wal, is a portmanteau of folks and taxonomy that particularly refers to indexing systems produced within internet communities. According to Vander Wal "a folksonomy is the result of personal free tagging of information for one's own retrieval. The tagging is done in a social environment by the person consuming the information." [Vander Wal, 2007]

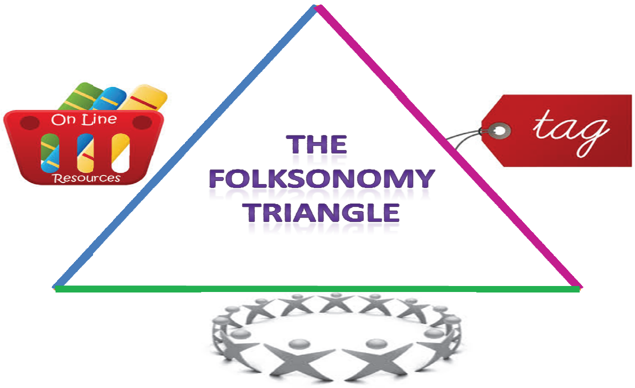


Figure 2.3: The folksonomy triangle: users, tags and resources

Folksonomies consist of three fundamental entities: users, tags, and resource. Users create tags to annotate resources such as: web pages, images, videos, etc. These tags are used to manage, index and classify online resources. Folksonomies also allow using these tags as a way to facilitate searches and navigate resources.

Types of folksonomies

There are two types of folksonomies. There are broad folksonomies which have several users who contribute in the creation of tags and narrow folksonomies where just a small number of users are tagging particular resources. A broad folksonomy allows many users to tag the same resources and any user can tag a resource using their own vocabulary. In a narrow folksonomy, only a few people are able to create tags and these tags are used by other users to locate resources. Unlike broad folksonomies, narrow folksonomies are not very common. An example of a broad folksonomy is del.icio.us, this is a website where users can tag any online resource they find relevant with their own personal tags. An example of a narrow folksonomies can be found in systems used by large businesses; these types of folksonomy are mainly used for research and associates working together in collaborative groups.

Strengths

There are key advantages that are important to understanding the utility of such systems.

a) **Browsing vs. Finding**

The first is serendipity. While the controlled vocabulary issues can hinder findability, browsing folksonomies and its interlinked tag sets is magnificent for finding new and unexpected content.

There is a basic difference in the activities of browsing to find interesting resources, as opposed to direct searching to find relevant documents in a query. Information seeking behavior varies based on context. While one could evaluate a folksonomy in a system by using specific queries from users, and then evaluating which documents tagged with keywords they choose are relevant to the query, which would ignore the broader set of browsing activities that the system seems to be stronger in. Measuring the utility of that aspect would likely require qualitative research in the form of interviews or ethnographic study of users, and is an area of further study. It would also require comparisons not to search based information retrieval systems, but to browsing activities using other categorization and classification schemes.

b) Users vocabulary

Maybe the most significant strength of a folksonomy is that it directly reflects the vocabulary of users. In an information retrieval system, there are at least two, and possibly many more vocabularies present [Buckland, 1999]. These could include that of the user of the system, the designer of the system, the author of the material, the creators of the classification scheme; translating between these vocabularies is often a difficult and defining issue in information systems. As discussed earlier, a folksonomy represents a fundamental shift in that it is derived not from professionals or content creators, but from the users of information and documents. In this way, it directly reflects their choices in diction, terminology, and precision.

Despite the power of web IR, its combination with social web it will give more improvement in term of effectiveness. In the following section a description of social information retrieval will be presented.

2.3 Social Information Retrieval

Based on the considerations presented above, this section presents a view about social web IR systems. The main motivation is to leverage the social network information for a deeper understanding of the relation between documents, queries and individuals to improve the relevance of retrieved documents to the user. The understanding of these relations is especially valuable in cases where a large number of relevant documents exist and insufficient information about documents or user's information needs are available. For the latter in particular, it has been shown that users in web IR systems tend to under specify their queries [Pasca, 2007; Silverstein et al., 1999]. In these cases, IR systems do not perform optimal in terms of retrieving relevant documents to the user. Using relations between documents greatly improves the ranking of documents under such conditions. However, the growing amount of information available on the web demands a constant improvement of relevance ranking algorithms. Therefore, the social web IR system may improve the relevance ranking by leveraging the social network extracted from the web documents. This can be summarized in following definition:

A social web IR system can be defined as a web IR system that integrates social network information into the information retrieval process.

This integration of social network information within a social web IR system is possible on different levels with regard to different intended goals for the selection of relevant documents.

2.3.1 Model for social information retrieval

Social information retrieval systems are distinguished from other types of IR systems by the incorporation of information about social relationships into the information retrieval process. This feature necessitates an extended model for information retrieval, as well as new techniques that make use of social information.

The traditional models for information retrieval concern themselves with documents, queries, and their relations to each other: A document is relevant to a query, a document references other documents, a query is similar to other queries, etc. Likewise, social network analysis models individuals and their relations with each other. Information retrieval systems traditionally do not model individuals, neither in their role as users of the system, nor as authors of the retrieved documents, and social networks do not incorporate retrievable content.

Social IR combines the models of information retrieval and social networks with each other. By incorporating individuals into the model, we gain a greater insight into their role in the information retrieval and production process (Figure 2.4).

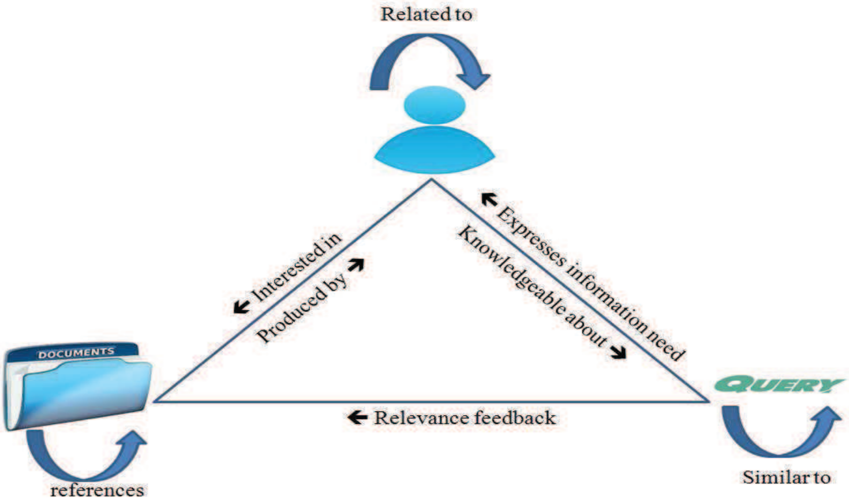


Figure 2.4: A model for social information retrieval

New associations between the entities become apparent: Individuals appear in their role as information producers or information consumers, queries relate to an individual's information needs, or describe a topic about which an individual possesses knowledge.

A social IR system is characterized by the presence of all three types of entities: documents, queries, and individuals. Most systems will only use a subset of the possible associations between the entities, depending on the domain of the system. Modeling the relations between individuals is mandatory for a social IR system; all other types of associations are optional, as long as all three entities have an association with at least one other.

Subsets of the web provide more suitable domains. The entirety of blog sites on the web is one such domain: Blog entries can usually be associated with an author, and via comments, communication between blog authors can be ascertained, leading to a social network. Wikis are also an environment that allows ascertaining authorship of a document, usually via the revision history. Interaction between users can be determined by co-authorship, or by discussions on dedicated talk pages; however, this information is often not portable between different wikis. Direct access to the underlying database often makes extraction of this information much easier.

Traditional information retrieval techniques which are based solely on analyzing document content, while very successful in many contexts, fail badly when the information need is underspecified, and when a large number of relevant documents exist. In this sense, social IR can be understood as a formalization of search techniques we commonly use to assess the quality of information by looking at the author's standing in his community.

Social IR aims at providing relevant content and information to users in the areas of information retrieval, and research; covering topics such as social tagging, collaborative querying, social network analysis, subjective relevance judgments, and collaborative filtering [Goh and Foo, 2007].

Several existing platforms investigate this track in order to improve the search paradigm. These platforms include Social Bing, Google+, etc. The research in this field has emerged and became very present in the daily life of users. Investigating the IR field from this perspective seems to be very promising to improve the representation, the storage, the organization, and the access to information.

2.4 Conclusion

In this chapter, we presented the fundamental concepts we are using. We introduced the notion of: (i) Information Retrieval by presenting its basic process, (ii) Social web by presenting and defining the main models of social relationships, and (iii) Social Information Retrieval by defining it as the concept that bridges the gap between IR and social web.

The next chapter is devoted to the state of the art of the most popular social information retrieval applications.

Chapter 3:
Social Information Retrieval: State of the Art

Chapter 3

Social Information Retrieval: State of the Art

There are two information access paradigms that users undertake each time they need to meet particular information needs on the web: *searching by query* and *recommendation*.

Querying a search engine is an effective approach that directly retrieves documents from an index of millions of documents in a fraction of a second. Details about this basic approach on social web, including collaborative tagging, personalization and ranking will be presented in the first section of this chapter (section 3.1). The other dominant information access paradigm involves recommendation-based systems that suggest items, such as movies, music or products by analyzing what the users with similar tastes have chosen in the past. Section 3.2 of this chapter details the main concepts and challenges related to social recommender systems.

3.1 Social web search

Social search is a kind of web search that takes into account the social aspect of the person initiating the search query. Results generated by social search engine give more visibility to content created or tagged by users in the social community. In this section, we survey several existing studies on collaborative tagging, personalized searches and ranking in social search processes.

3.1.1Collaborative Tagging and Searching

Social tagging is the activity of annotating digital resources with keywords, so-called tags [Golder and Huberman, 2006; Trant, 2009]. In social tagging systems, users can annotate a variety of digital resources with tags, for instance, bookmarks, pictures, or products. In most applications, users are free to choose any tags for describing their resources in order to structure and organize their own stored web material. The tags which are used will reflect individual associations with regard to resources, and they will describe a specific meaning or relevance for the respective users.

The social aspect of social tagging systems lies in the opportunity to use other people's tags as navigation links for one's own search processes. Social tagging systems aggregate the tags of all users and describe the resources in a so-called folksonomy [Trant, 2009; Vander Wal, 2005]. The tags of many different users are aggregated and the resulting collective tag structure depicts the collective knowledge of web users. The individual users' tags establish a network of connections between resources and tags, and among those tags themselves. The more frequently tags are used for one resource; the stronger becomes the connection among them. Analogously, the more often two tags co-occur for one resource, the stronger they are related to each other. When aggregating all tags from a community, a collective representation of the connections between related tags and their strengths of association will emerge. These associations are typically visualized by tag clouds, in which different font sizes represent the strength of association of tags to a related tag or a resource.

Social tagging systems augment the collective structure of a community with the individual knowledge representations of individual users. Tag clouds externalize the community's associations between tags and the strengths of associations. In this way, social tags are able to provide visual representations of the conceptual structure of a domain, which is built upon the knowledge of individuals who belong to a large web community. In this section we outline the principle of collaborative information retrieval, then we address the principal challenges in social tagging activities, and next we survey works that tackled these defies in social information retrieval.

Collaborative information retrieval

Social platforms allow users to provide, publish and spread information, like commenting or tweeting about an event. In such a context, a huge quantity of information is created in social media, which represents a valuable source of relevant information. Hence, many users use social media to gather recent information about a particular content by searching collection of posts and status. Therefore, social content search systems come as a mean to index content explicitly created by users on social media and provide a real-time search support [Jansen et al., 2010].

There are several social content search engines, which index real-time content spreading systems. This includes TwitterSearch, Social Bing, etc. Social content search systems deal with a different kind of content than classic search engines. Indeed, posts and statuses published on social media are often short, frequent, and do not change after being published, while web pages are rich, generated more slowly, and evolve after creation [Teevan et al., 2011]. Dealing with such content is challenging, because it requires real-time and recency sensitive queries processing.

Sensitive query refers to a query where the user expects documents, which are both topically relevant as well as fresh [Dong et al., 2010]. A study has been performed by [Teevan et al., 2011] that give an overview of "What is the motivation behind a user to use a social content search system rather than a classic search engine?". This study reveals that social content search systems are interrogated with queries, which are shorter and more popular. The main goal is to find temporally relevant information and information related to people. One of the weaknesses of search engines available today (e.g. Google, Yahoo!, Bing) is the fact that they are designed for a single user who searches alone. Thus, users cannot benefit from the experience of each other for a given search task. [Morris, 2007, 2008] conducts a survey on 204 knowledge workers in a large technology company in which she revealed that 97% of respondents reported engaging in one of collaborative search task described in the survey. For example, 87,7% of respondents reported having watched over someone's shoulder for query suggestion, and 86,3% of respondents reported having e-mailed someone to share the results of a web search.

In such a context, [Morris and Horvitz, 2007] developed SearchTogether, a collaborative search interface, where several users who share an information need collaborate and work together with others to fulfill that need. The authors discuss the way SearchTogether facilitates collaboration by satisfying criteria like awareness, division of labor, and persistence. Similarly, [Filho et al., 2010] proposed Kolline, a search interface that aims at facilitating information seeking for inexperienced users by allowing more experienced users to collaborate together.

[Paul and Morris, 2009] investigate sense-making for collaborative web search, which is defined as the act of understanding information. The study revealed several themes regarding the sense-making challenges of collaborative web search, e.g. Awareness, Timeliness and sense-making hand-off. Based on their finding, they proposed CoSense, a system that supports sense-making for collaborative web search tasks that provides enhanced group awareness by including a time-line view of all queries executed during the search process. Even though these features help to enhance participants' communication and sense-making during their search activities, users still have to sort among different documents and analyze them one by one to find relevant information.

Because we are interested in chapter 5 by information retrieval in collaborative E-learning; an overview about the main contributions related to this field will be presented.

In [Westerski, et al., 2006], the authors proposed an approach based on gathering all student interactions and activities to exploit it on data about student's current course progression. In [Torniai et al., 2008], the authors tried to leverage the social semantic web paradigm where they proposed a collaborative semantic-rich learning environment in which folksonomies are created from students' collaborative tags.

In the work of [Wong et al., 2013], the authors presented findings from a small scale study exploring the first year students' experiences on social networking usage and their perception on using it for e-learning. The data are collected quantitatively, consisting of surveys on students' experiences on using social networking as a communication and collaboration tool.

The approach of [Mutschke and Mayr, 2015] studied the applicability and usefulness of two particular science models for re-ranking search results (Bradfordizing and author centrality). The authors provides a preliminary evaluation study that demonstrates the benefits of using science model driven ranking techniques, but also how different the quality of search results can be if different conceptualizations of science are used for ranking.

The contribution of [Shi et al., 2013] introduced Topolor, a social personalized adaptive e-learning system, which aims at improving fine-grained social interaction in the learning process in addition to applying classical adaptation based on user modeling.

The work of [Boff and Reategui, 2012] presents a learning environment where a mining algorithm is used to learn patterns of interaction with the user and to represent these patterns in a scheme called item descriptors. The learning environment keeps theoretical information about subjects, as well as tools and exercises where the student can put into practice the knowledge obtained. The students' actions, as well as their interactions, are monitored by the system and used to find patterns that can guide the search for students that may play the role of a tutor.

In the contribution of [Chan and Jin, 2006], the authors proposed a collaboratively shared Information Retrieval model to complement the conventional IR approach (i.e. objective) with the collaborative user contribution (i.e. subjective). The proposed architecture and mechanisms provide a way to handle general purpose textual information and provide advanced access control features to the system.

The objective of [Moutachaouik et al., 2012] work is the conception and the realization of a recommendatory system, using concepts of the web usage mining and being inspired by approaches to information filtering. This system includes a new hybrid method to rank documents web, in order to propose to the webmaster (or admin) of platform e- learning the best available documents based of the historical to research done by learners. The elaborated system will make it possible to propose help and assistance to learners of the system.

According to [Wang et al., 2011], learning efficiency can be greatly improved if E-learning users' social networks properties can be effectively utilized. The focus of the authors' study is on E-learners' positive influence between their relationships, where the authors proposed a new model and selection algorithm named Weight Positive Influence Dominating Set (WPIDS) and analyzed its efficiency through a case study.

In the work of [Shi et al., 2014], the authors proposed a set of contextual strategies, which apply flow and self-determination theory for increasing intrinsic motivation in social e-learning environments. The authors also present a social e-learning environment that applies these strategies, followed by a user case study, which indicates increased learners' perceived intrinsic motivation.

In their work, [Chavarriaga et al., 2014] proposed a system to recommend activities and resources that help students in achieving competence levels throughout an online or blended course. This recommender system takes into consideration experiences previously stored and ranked by former students.

[Palazuelos et al., 2013] exposed how social network analysis can be a tool of considerable utility in the educational context for addressing difficult problems, e.g., uncovering the students' level of cohesion, their degree of participation in forums, or the identification of the most influential ones.

[Köck and Paramythis, 2011] present an approach based on the modeling of learners' problem solving activity sequences, and on the use of the models in targeted, and ultimately automated clustering, resulting in the discovery of new, semantically meaningful information about the learners. The approach is applicable at different levels: to detect pre-defined, well-established problem solving styles, to identify problem solving styles by analyzing learner behavior along known learning dimensions, and to semi-automatically discover learning dimensions and concrete problem solving patterns.

In [Li and Iribe, 2012], the authors described a practical use of social networking service to support continued communication among adult students in an e-learning program.

To position our contribution in collaborative E-learning, we will propose, in chapter 5, a new approach for personalize the results proposed to each learner when he seeks to find relevant resources by using tags.

In another side, event detection in social web has recognized a real success in the last years especially in information retrieval processes. For example, in the contribution of [Sakaki et al., 2010] the authors investigated the real-time interaction of events such as earthquakes, in

Twitter, and proposed an algorithm to monitor tweets to detect a target event. In another contribution, [Becker et al., 2011] explored approaches to analyze the stream of Twitter messages to distinguish between messages about real-world events and non-event messages. The approach relies on a rich family of aggregate statistics of topically similar message clusters. The main contribution of [Wang et al., 2012] is to incorporate online social interaction features in the detection of physical events. In another contribution, [Abdelhaq et al., 2013] presented a framework to detect localized events in real-time from a Twitter stream and to track the evolution of such events over time. To determine the most important events in a recent time frame, the authors introduced a scoring scheme for events. [Psallidas et al., 2013] discussed both the unknown and known event identification scenarios, and attempted to characterize the key factors in the identification process. Furthermore, the authors proposed novel features of the social media content to exploit it as well as the modeling of the typical time decay of event-related content. The approach of [Alonso and Shiells, 2013] based on constructing a particular game's timeline in such a way that it can be used as a quick summary of the main events that happened along with popular subjective and opinionated items that the public inject. In this work, the authors introduced a timeline design that captures a more complete story of the event by placing the volume of Twitter posts alongside keywords that are driving the additional traffic.

In folksonomies a relatively large number of resources can match users' queries. Therefore, ranking these web resources is a key problem, since a user cannot browse all them. Consequently, in this thesis we want to benefit from the dependence between the presence of an event and the high number of similar queries transmitted in the same period to improve resources retrieval and ranking in collaborative applications. Details will be presented in the following chapter.

Challenges and limitations

The Web 2.0 consists essentially in a successful evolution of traditional web applications supported by some principles and technologies. Social tagging and the resulting folksonomies can be seen as two of those principles that have emerged and met a big success within social web applications. The simplicity of tagging combined with the culture of exchange allows the mass of users to share their annotations on the mass of resources. However, the exploitation of folksonomies raises several issues highlighted by [Mathes, 2004] and by [Passant, 2009]:

a) Tag ambiguity (Polysemy)

Polysemy refers to a word that has two or more meanings. "Poly" means 'many', and "semy" means 'meanings'. A polysemous word is one that has many ("poly") senses ("semy").

As an uncontrolled vocabulary that is shared across the entire system, the terms in a folksonomy have inherent ambiguity as different users apply terms to documents in different ways. There are no explicit systematic guidelines and no scope notes.

For example, consider the following two assignments: <u1, “apple”, r1>, <u2, “apple”, r2> (users u1 and u2 use “apple” to tag two different resources r1 and r2).

Even if u1 and u2 use “apple” to express different ideas, the system would still return both r1 and r2 when another user (who might have a totally different idea of “apple”) searches for “apple”.

b) Acronyms

Acronyms present another area of potential ambiguity. For example examining the front page of del.icio.us on April 12, 2015 revealed one user tagging sites with “ANT.” After examining the other sites the user tagged with ANT, it was apparent this was an acronym for “Actor Network Theory,” in the domain of sociology. However, when examining the ANT tag across all users most of the bookmarks were about Apache Ant, a project building tool in the Java programming language. Two completely separate domains and ideas are mixed together in the same tag.

c) Synonyms (Spelling variations)

There is no synonym control in tagging systems. This leads to tags that seemingly have similar intended meanings. Consider the following two assignments: <u1, “cat”, r1>, <u2, “kitty”, r2> (users u1 and u2 use “cat” and “kitty” to tag two different resources r1 and r2, respectively.). Even if u1 and u2 have the same meaning in mind for the two different tags, the system is not able to relate these tags in a general sense.

Different word forms, plural and singular, are also often both present. In the popular tags on Flickr, both plurals and singulars were listed.

d) The lack of semantic relations between tags

The lack of explicit representations of the knowledge contained in folksonomies where the semantic relations that may exist between tags are not represented, as for example with the tags “car” and “vehicle” where it is possible to state that a “car” “is a type” of “vehicle”.

e) Dealing with different languages

The difficulties to deal with tags from different languages, since this information is generally not provided at tagging time, and several languages can be mixed in an open web platform, and even for an individual user who uses several languages to communicate. This problem concerns the lack of explicit specification of the language of a given tag, which can raise issues when attempting to structure tags. For example, if several languages are used to tag a given resource, it is hard to guess whether some tags are translation of other tags or different concepts.

These sorts of problems are the reasons why controlled vocabularies are used in many settings. Generally, any of the classic problems that controlled vocabularies help deal with will be present in these systems to varying degrees. However, it is likely that a controlled vocabulary would be impossible in the context of systems like Delicious and Flickr.

Extracting the semantics of folksonomies

In this section, we focus on methodologies and systems aimed at uncovering the emergent semantics from folksonomies. Since usually no explicit semantic relationships are given when users tag, tag semantics have to be captured by overcoming tags ambiguity and spelling variations.

a) Resolving Tag Ambiguity

Among the most important contributions on resolving tag ambiguity and extracting the semantic links between tags in a folksonomy, we start with [Mika, 2005] who has proposed to extend the traditional bipartite model of ontologies to a tripartite model. In his contribution, Mika focuses on social network analysis in order to extract lightweight ontologies, and therefore semantics between the terms used by the actors. [Gruber, 2005] recommended to build an ontology of folksonomy. According to him, the problem of the lack of semantic links between terms in folksonomies can be easily resolved by representing folksonomies with ontologies. [Specia and Motta, 2007] proposed a method consisting in building clusters of tags, and then trying to identify possible relationships between tags in the same cluster. The authors have chosen to reuse available ontologies in order to represent the correlations which hold between tags. An attempt to automate this method has been done by [Angeletou et al., 2007].

[Buffa et al., 2008] present a semantic wiki with the aim of exploiting the force of ontologies and semantic web standard languages in order to improve social tagging. According to the authors, with this approach, tagging remains easy and becomes both motivating and unambiguous. The niceTag project of [Limpens et al., 2010] is focused on using ontologies to extract semantics between tags in a system. In addition, the interactions among users and the system are used to validate or invalidate automatic treatments carried out on tags. The authors have proposed methods to build lightweight ontologies which can be used to suggest terms semantically close during a tag-based search of documents. [Pan et al., 2010] addressed the tag ambiguity problem by extending folksonomy with ontologies. They proposed to expand folksonomies in order to avoid bothering users with the rigidity of ontologies. During a keyword-based search of resources, the set of ambiguous used terms is concatenated with other tags so as to increase the precision of the search results. [Wu and Zhou, 2011] tried to estimate the semantic relations among tags to judge if tags are related from semantic view or isolated. The authors proposed to perform several measures of semantic relatedness to discover semantic information within a folksonomy.

To sum up, most of the works aspire to bring together ontologies and folksonomies as a solution to resolve tag ambiguity and overcome the lack of semantic links between tags. Sure enough the approaches described in this section show that the social nature of resource sharing is not in contradiction with the possibilities offered by ontology-based systems. But the rigidity that characterizes ontologies and the need for an expert who must control and organize the links between terms as in [Gruber, 2005] seem a little cumbersome and too much expensive. Even the structures automatically extracted as in [Mika, 2005] still suffer from the ambiguity of concepts. Regarding the work of [Specia and Motta, 2007], we can say that the use of semantic web ontologies for extracting relationships between terms is not sufficient, because as the semantic web includes some specific domain ontology, that will push back the problem. Also the expertise of users which was introduced in [Limpens et al., 2010] is characterized by the complexity of its exploitation. As a result, in chapter 4 we propose an approach of tag-based resource recommendation where we aim to resolve tag ambiguity and spelling variations without explicitly using ontologies. We base upon association rules which are a powerful method to discovering interesting relationships among a large dataset on the web. Our aim is to enrich user profiles based on similarities between users and association rules and by doing so to increase the community effect when suggesting resources to a given user.

b) Dealing with spelling variations problem

The goal here is to detect and group tags that are equivalent in their meanings or in the topic they describe but are spelled with some variations, such as in “math” and “mathematic” for example. In this part, we do not consider the structure of folksonomies but simply focus on the morphological similarity of tags two by two. The main types of methods are the following:

-String-based methods: this type of method measure the difference between the strings of characters of the tags. It has been used, for instance, by [Specia and Motta, 2007] to group spelling variants tags.

-Linguistic methods: These methods seek to exploit some linguistic or semantic properties of the words to draw comparison between them. For instance, stemming algorithms consist in extracting roots from words (e.g. "links" and "linked" become "link") and grouping tags sharing the same roots. It is also possible to exploit additional resources. For example, [Specia and Motta, 2007; Van Damme et al., 2007] suggest using online resources (such Wikipedia, or online dictionaries) to check the correct spelling of tags or to find an appropriate representative for a cluster of equivalent tags (grouped together thanks to string-based method for instance).

[Euzenat and Shvaiko, 2007] also give a detailed overview of these two types of methods when they utilized for matching similar concepts from different ontologies.

A first distinction among the different metrics to be used to compare tag labels is the difference between distance functions and similarity functions. Distance functions associate a real number *d* to a pair of strings *(s1,s2)*, where the smaller the value of *d*, the closer the strings. Similarity functions associate a real number *o* to a pair of strings *(s1,s2)*, where the greater the value of *o*, the closer the strings. In the SimMetrics[1] package, all measures are implemented so that they can be considered as similarity metrics, even though they can make use of distances, like edit distances, to compute a similarity.

The similarity metrics of this package fall into several categories: (a) edit distance based methods, which consider the set of operations needed to turn string s1 into string s2, such as e.g. Levenshtein, or Gotho; (b) token-based methods, which decompose strings into sets of substrings such as Overlap Coefficient or Monge-Elkan ; (c) token-based methods using vector representations of strings such as the cosine similarity; and finally (d) other types of metrics such as QGram or Soundex metrics that compare different features of strings (Soundex e.g. associates an arbitrary code to letters composing a string so that string that sound similar have the same code, as e.g. "robert" and "rupert").

[1] http://www.dcs.shef.ac.uk/~sam/stringmetrics.html

A simple way to detect equivalent tags using these distance metrics consists in choosing a threshold value above which two tags are considered equivalent.

Although the methods described in the above-mentioned works are able to deal with spelling variation, the main disadvantage of those approaches is don't consider the social structure of folksonomies but merely focus on the morphological similarity of tags two by two. In our contribution, we present a method to treat spelling variations problem based on social relation between users.

3.1.2 Personalization

Recently, several search tools for the web have been developed to tackle the information overload problem. Some make use of effective personalization, adapting the results according to each user's information needs. This contrasts with traditional search engines that return the same result list for the same query, regardless of who submitted the query, in spite of the fact that different users usually have different needs.

In the last few years, attention has focused on the adaptation of traditional IR system to the web environment, and related implementations of personalization techniques. For this reason, sophisticated search techniques are required, enabling search engines to operate more accurately for the specific user, abandoning the "one-size-fits-all" method.

Personalized search aims to build systems that provide individualized collections of pages to the user, based on some form of model representing their needs and the context of their activities. Depending on the searcher, one topic will be more relevant than others. Given a particular need, e.g., a query, the results are tailored to the preferences, tastes, backgrounds and knowledge of the user who expressed it.

This section provides a brief overview of the personalized search approaches with a broad description of the various methods and techniques proposed in the literature. We begin with a general idea about content and collaborative-based distinction. We then move on to how user profiles are implemented in the personalized systems and the typical sources employed to recognize user needs. An overview of the different personalized search approaches in collaborative tagging systems closes this section.

Content and Collaborative-Based Personalization

The content of documents is used to build a particular representation that is exploited by the system to suggest results to the user in response to queries. The searching by query paradigm is definitely quicker when the user is aware of the problem domain and knows the appropriate discerning words to type in the query [Olston and Chi, 2003]. However, analyzing search behavior, it is possible to see that many users are not able to accurately express their needs in exact query terms. The average query contains only 2 to 3 terms [Lawrence and Giles, 1995; Spink and Jansen, 2004].

Due to *polysemy*, the existence of multiple meanings for a single word, and *synonymy*, for the existence of multiple words with the same meaning, the keyword search approach suffers from the so-called *vocabulary problem* [Furnas et al., 1987]. This phenomenon causes mismatches between the query space and the document space, because a few keywords are unlikely to select the right resources to retrieve from sets of billions [Freyne and Smyth, 2004]. Synonymy causes relevant information to be missed if the query does not contain the exact keywords occurring in the documents, inducing a recall reduction. Polysemy causes irrelevant documents to appear in the result lists, affecting negatively the system precision.

For these reasons, users face a difficult battle when searching for the exact documents and products that match their needs. Understanding the meaning of web content and, more importantly, how it relates to the real meaning of the user's query, is a crucial step in the retrieval process.

When the algorithm used to build the result list also takes into account models of different users, the approach is usually named *collaborative* [Goldberg et al., 1992; Resnick et al., 1994]. The basic idea behind collaborative-based approaches is that users with similar interests are likely to find the same resources interesting for similar information needs. *Social navigation* is the word coined by [Dieberger et al., 2000] to refer to software that allows people to leave useful traces on web sites, such as reviews, comments, or votes, used by other people during browsing and searching-by-query.

User Modeling in Personalized Systems

Tracking what pages the user has chosen to visit and their submitted queries is a type of *user modeling* or *profiling* technique, from which important features of users are learned and then used to get more relevant information.

In the simplest cases, user models consist of a registration form or a questionnaire, with an explicit declaration of interest by the user. In more complex and extended cases, a user model consists of dynamic information structures that take into account background information, such as educational level and the familiarity with the area of interest, or how the user behaves over time. In personalized search systems the user modeling component can affect the search in three distinct phases:

– *Part of retrieval process*: the ranking is a unified process wherein user profiles are employed to score web contents.

– *Re-ranking*: user profiles take part in a second step, after evaluating the corpus ranked via non-personalized scores.

– *Query modification*: user profiles affect the submitted representation of the information needs, e.g., query, modifying or augmenting it.

The first technique is more likely to provide quick query response, because the traditional ranking system can be directly adapted to include personalization, avoiding repeated or superfluous computation. However, since the personalization process usually takes a long time compared with traditional non-personalized IR techniques, most search engines do not employ any personalization at all. Time constraints that force the system to provide result lists in less than a second cannot be met for all users.

On the other hand, re-ranking documents allows the user to selectively employ personalization approaches able to increase precision. Many systems implement this approach on the client-side, e.g., [Pitkow et al., 2002; Micarelli and Sciarrone, 2004; Speretta and Gauch, 2005], where the software connects to a search engine, retrieving query results that are then analyzed locally. In order to avoid spending time downloading each document that appears in the result list, the analysis is usually only applied to the top ranked resources in the list, or it considers only the snippets associated with each result returned by the search engine.

Because of the time needed to access a search engine and retrieve the resources to be evaluated, the re-ranking approach implemented via client-side software can be considerably slow. Nevertheless, complex representations of user needs can be employed, considerably improving the personalization performances.

Finally, profiles can modify the representations of the user needs before that retrieval takes place. For instance, if the user needs are represented by queries, the profile may transform them by adding or changing some keywords to better represent the needs in the current profile. Short queries can be augmented with additional words in order to reduce the vocabulary problem, namely, polysemy and synonymy, which often occur in this kind of keyword-based interaction. Alternatively, if the query retrieves a small number of resources, it is possible to expand it using words or phrases with a similar meaning or some other statistical relations to the set of relevant documents. The major advantage of this approach is that the amount of work required to retrieve the results is the same as in the unpersonalized scenarios. Nevertheless, user profiles affect the ranking only by altering the query representations. Unlike ranking that takes place in the retrieval process, the query modification approach is less likely to affect the result lists, because it does not have access to all the ranking process and its internal structures.

Sources of Personalization

The acquisition of user knowledge and preferences is one of the most important problems to be tackled in order to provide effective personalized assistance. Some approaches employ data mining techniques on browsing histories or search engine logs, while others use machine learning [Webb et al., 2001] to analyze user data, that is, information about personal characteristics of the user, in order to learn the knowledge needed to provide effective assistance. The user data usually differs from usage data. The latter are related to a user's behavior while interacting with the system. Examples of sources of user data are: personal data, e.g., name, address, phone number, age, sex, education; or geographic data, e.g., city and country.

Techniques such as *relevance feedback* and *query expansion* introduced in the IR field [Salton and McGill, 1983; Allan, 1996] can be employed in the personalization domain in order to update the profile created by users. Basically, to improve ranking quality, the system automatically expands the user query with certain words that bring relevant documents not literally matching the original query. These words are usually extracted from resources in a previously retrieved list of ranked documents that have been explicitly judged interesting by the user through relevance feedback.

Besides considering important synonyms of the original queries' keywords that are able to retrieve additional documents, expansion helps users to disambiguate queries.

For example, if the user submits the query '*Jaguar*', the result list will include information on the animal, the car manufacturer, the operating system, etc. Following relevance feedback on a subset of documents relating to the meaning of interest to the user, the query is updated with words that help the system filtering out the irrelevant pages. Using a lexicon, it is also possible to expand queries such as '*IR*' to '*information retrieval*', increasing the chance of retrieving useful pages.

Even though these techniques have been shown to improve retrieval performance, some studies have found that explicit relevance feedback is not able to considerably improve the user model especially if a good interface is not provided to manage the model and clearly represent the contained information [Wærn, 2004]. Users are usually unwilling to spend extra effort to explicitly specify their needs or refine them by means of feedback [Anick, 2003], and they are often not able to use those techniques effectively [Spink, 2000; Teevan et al., 2005], or they find them confusing and unpredictable [Koenemann and Belkin, 1996].

Moreover, studies show that users often start browsing from pages identified by less precise but more easily constructed queries, instead of spending time to fully specify their search goals [Teevan et al., 2004]. Aside from requiring additional time during the seeking processes, the burden on the users is high and the benefits are not always clear, therefore the effectiveness of explicit techniques may be limited.

Because users typically do not understand how the matching process works, the information they provide is likely to miss the best query keywords, i.e., the words that identify documents meeting their information needs. Moreover, part of the user's available time must be employed for secondary tasks that do not coincide with their main goal. Instead of requiring user's needs to be explicitly specified by queries or manually updated by the user feedback, an alternative approach to personalize search results is to develop algorithms that infer those needs implicitly.

Basically, *implicit feedback* techniques unobtrusively draw usage data by tracking and monitoring user behavior without an explicit involvement. Personalized systems can collect usage data on the server-side, e.g., server access logs or query and browsing histories, and/or on the client-side, such as cookies and mouse/keyboard tracking.

Personalized searches in collaborative tagging systems

With the recent development of collaborative tagging systems, we have begun to see some works proposed for personalized searches in the collaborative tagging systems. [Noll and Meinel, 2007] proposed a simple effective approach to explore user and resource-related tags based on TF. These authors re-ranked the non-personalized search results based on these related tags. [Xu et al., 2008] proposed topic-based personalized searches in folksonomy in which the personalized search is conducted by ranking the resources based not only on term similarity matching but also on topic similarity matching. Instead of using TF in their work, [Xu et al., 2008] used TF-IDF and BM25 (BM stands for Best Matching) to construct user and resource profiles.

As a follow-on study to [Xu et al., 2008], [Vallet et al., 2010] used different techniques to measure the user-resource similarities and compare the effect of these techniques.

Although the methods described in the above-mentioned work are able to handle personalized searches with tag-based user and item profiles, there are some limitations. These limitations are examined in following chapters where our contribution will be detailed.

3.1.3 Ranking

When the user gives a query, the index is consulted to get the documents most relevant to the query. The relevant documents are then ranked according to their degree of relevance, importance, etc. In IR, ranking results consists in the definition of a ranking function that allows quantifying the similarities among documents and queries. We distinguish two categories for social results ranking that differ in the way they use social information.

The first category uses social information by adding a social relevance to the ranking process, while the second uses it to personalize search results.

Ranking using social relevance

Several approaches have been proposed to improve document ranking using social relevance. Social relevance refers to information socially created that characterizes a document from a point of view of interest, i.e. its general interest, its popularity, etc. Two formal models for folksonomies and ranking algorithm called folkRank and SocialPageRank have been proposed in [Hotho et al., 2006] and [Bao et al., 2007] respectively. Both are an extension of the well-known PageRank algorithm adapted for the generation of rankings of entities within folksonomies. SocialPageRank intends to compute the importance of documents according to the mutual enhancement relation among popular resources, up-to-date users and hot social annotations. In the same spirit, [Takahashi et al., 2008, 2009] propose S-BIT (Social-Bookmarking Induced Topic Search) and FS-BIT (Freshness Social-Bookmarking Induced Topic Search), extensions of the well-known HITS (Hyperlink-Induced Topic Search) [Kleinberg, 1999] approach. Finally, [Yanbe et al., 2007] proposed SBRank (Social Bookmarking Rank), which indicates how many users bookmarked a page, and use the estimation of SBRank as an indicator of Web search. All these algorithms are in the context of folksonomies, and a number of them are reviewed and evaluated in [Abel et al., 2008].

Personalized ranking

In general, users have different interests, different profiles, and different habits. Consequently, in an IR system, providing the same documents sorted in the same way is not really suitable. Thus, a personalized function to sort documents differently according to each user is expected to improve search results.

Several approaches have been proposed to personalize ranking of search results using social information [Bender et al., 2008; Carmel et al., 2009; Noll et al., 2009; Vallet et al., 2010; Wang and Jin, 2010 and Xu et al., 2008]. Almost all these approaches are in the context of folksonomies and follow the same idea that the ranking score of a document *d* retrieved when a user u submits a query q is driven by: (i) a term matching process, which calculates the similarity between q and the textual content of d to generate a user unrelated ranking score; and (ii) an interest matching process, which calculates the similarity between u and d to generate a user related ranking score. Then a merge operation is performed to generate a final ranking score based on the two previous ranking scores.

Because we are interested with personalized ranking functions based on folksonomies, in this section, we formally define some different personalized ranking functions. We each time present the ranking score of a document d for a query q issued by a user u denoted *Rank (d, q, u)*.

Profile Based Personalization: The approach presented by [Xu et al., 2008] assumes the ranking score of a document *d* is decided by two aspects: (i) a textual matching between *q* and *d*, and (ii) a user interest matching between *u* and *d*. Hence, their approach can be defined as follows:

$$Rank(d,q,u) = \gamma \times Cos(\overrightarrow{p_u}, \overrightarrow{T_d}) + (1-\gamma) \times Sim(\vec{q}, \vec{d}) \qquad (3.1)$$

Where, γ is a weight that satisfies $0 \leq \gamma \leq 1$ and $Sim(\vec{q}, \vec{d})$ denotes the textual matching score between *d* and q.

Topics Based Personalization (LDA-P): We present here a topics-based approach. This approach is based on Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. LDA-P relies on the fact that the set of tags can be used to represent web pages and as input for LDA to construct a model. Then, for each document that matches a query, LDA-P computes a similarity between its topic and the topic of the user profile using the cosine measure (inferred using the previous constructed LDA model). The obtained similarity value is merged with the textual ranking score to provide a final ranking score for a document that matches a query w.r.t the query issuer as follows:

$$Rank(d,q,u) = \gamma \times Cos(\overrightarrow{u_{topic}}, \overrightarrow{d_{topic}}) + (1-\gamma) \times Sim(\vec{q}, \vec{d}) \qquad (3.2)$$

Where, $0 \leq \gamma \leq 1$, $\overrightarrow{u_{topic}}$ and $\overrightarrow{d_{topic}}$ are respectively the vectors that model the user and the document topics based on the constructed LDA model.

Social Context Based Personalization (SoPRa): The approach proposed by [Bouadjenek et al., 2013] is similar to [Xu et al., 2008]. However, the authors proposed to enhance the ranking process by considering a new aspect, which is the social matching score. This approach takes into account the entire social context that surrounds both users and documents and is called SoPRa. It is defined as follows (β is set to 0.5):

$$Rank(d,q,u) = \gamma \times Cos(\overrightarrow{p_u}, \overrightarrow{T_d}) + (1-\gamma) \times [\beta \times Cos(\vec{q}, \overrightarrow{T_d}) + (1-\beta) \times Sim(\vec{q}, \vec{d})] \quad (3.3)$$

Scalar Tag Frequency Based Personalization: The approach presented by [Noll and Meinel, 2007] considers only a user interest matching between *u* and *d*. This approach does not make use of the user and document length normalization factors, and only uses the user tag frequency.

The authors normalize all document tag frequencies to 1, since they want to give more importance to the user profile. Their ranking function can be defined as follows:

$$Rank(d,q,u) = \sum_{t \in T_u \wedge t \in T_d} |D_{u,t}| \quad (3.4)$$

Scalar tf-if Based Personalization (tf-if): [Vallet et al., 2010] proposed to improve the [Noll and Meinel, 2007] approach above by including a weighting scheme based on an adaptation of the tf-idf as follows:

$$Rank(d,q,u) = \sum_{t \in T_u \wedge t \in T_d} (tf_u(t) \times iuf(t) \times tf_d(t) \times idf(t)) \quad (3.5)$$

In the following chapter, we propose a new ranking function based on social similarity and event detection. Motivations and details about this function will be described next.

3.2 Social Recommender Systems

Recommender systems have become an important research area since the appearance of the first contributions on collaborative filtering in the mid-1990s [Hill et al., 1995, Resnick et al., 1994, Shardanand and Maes, 1995].

There has been much work done on developing new approaches to recommender systems over the last decade. The interest in this domain still remains high because it constitutes a problem-rich research area and because of the abundance of practical applications that help users to deal with information overloads and provide personalized recommendations, content, and services to them.

Social tagging systems have grown in popularity over the web in the last years on account of their simplicity to categorize and retrieve content using tags. The increasing number of users providing information about themselves through social tagging activities caused the emergence of tag-based profiling approaches, which assume that users expose their preferences for certain contents through tag assignments. Thus, the tagging information can be used to make recommendations. After giving a general idea about traditional recommender systems, this section will present an overview about how can social tagging systems be used for extending the capabilities of recommender systems.

3.2.1 Traditional Recommender Systems

Recommender systems combine ideas from user profiling, information filtering and machine learning to deliver users a more intelligent and proactive information service by making concrete product or service recommendations that match their learned user preferences and needs. The recommender technology is superior to other information filtering applications because of its ability to provide personalized and meaningful information recommendations.

For example, while standard search engines are very likely to generate the same results to different users entering identical search queries, recommender systems are able to generate results to each user that are personalized and more relevant because they take into account each user's personal interests.

In general, two recommendation techniques have come to dominate: content-based filtering (CBF) and collaborative filtering (CF). The content-based approach [Mooney and Roy, 2000] recommends to a user items whose content is similar to content that the user has previously viewed or selected. In a movie recommender application, for instance, a CBF system will typically rely on information such as genre, actors, director, producer etc. and match this against the learned preferences of the user in order to select a set of promising movie recommendations.

CBF recommender systems need a technique to represent the features of the items. Feature representation can be created automatically for machine-parsable items (such as news or papers) but must be manually inserted by human editors for items that are not yet machine-parsable (such as movies and songs). Obviously this activity is expensive, time consuming, error-prone and highly subjective. Moreover, for some items such as jokes, it is almost impossible to define the right set of describing features and to "objectively" classify them [Massa and Bhattacharjee, 2004].

Collaborative filtering (CF) collects information about a user by asking them to rate items and makes recommendations based on highly rated items by users with similar taste. CF approaches make recommendations based on the ratings of items by a set of users (neighbors) whose rating profiles are most similar to that of the target user [Breese et al., 1998].

CF algorithms generally compute the overall similarity or correlation between users, and use that as a weight when making recommendations. In a book recommendation application, for example, the first step for the CF system is try to find the "neighbors" of the target user.

The “neighbors” refer to other users who have similar tastes in books (rate the same books similarly). In the second step, only the books that are highly rated by the “neighbors” would be recommended.

In contrast with the content-base approaches, the CF techniques rely on the availability of user profiles that capture the past ratings histories of users [Breese et al., 1998] and don’t require any human intervention for tagging content because item knowledge is not required.

Therefore, the CF techniques can be applied to virtually any kind of items: papers, news, web sites, movies, songs, books, jokes, locations of holidays, stocks and promise to scale well to large item bases [Massa and Bhattacharjee, 2004]. Collaborative filtering is the most widely used approach to build online recommender systems. It has been successfully employed in many applications, such as recommending books, CDs, and other products at Amazon.com, Movies by MovieLens [Adomavicius and Tuzhilin, 2005]. Some methods combine both content and collaborative filtering approaches to make recommendations [Schein et al., 2002].

3.2.2 Challenges and Limitations

In this subsection, we present some of the common difficulties in deploying recommender systems, as well as some research directions that addressed them.

Sparsity

Stated simply, most users do not rate most items and hence the user ratings matrix is typically very sparse. This is a problem for Collaborative Filtering systems, since it decreases the probability of finding a set of users with similar ratings. This problem often occurs when a system has a very high item-to-user ratio, or the system is in the initial stages of use. This issue can be mitigated by using additional domain information [Melville et al., 2002] or making assumptions about the data generation process that allows for high-quality imputation [Su et al., 2008].

The Cold-start Problem

New items and new users pose a significant challenge to recommender systems. Collectively these problems are referred to as the cold-start problem [Schein et al., 2002]. The first of these problems arises in Collaborative Filtering systems, where an item cannot be recommended unless some user has rated it before. This issue applies not only to new items, but also to obscure items, which is particularly detrimental to users with eclectic tastes. As such the new-item problem is also often referred to as the first-rater problem. Since content-based approaches [Mooney et al., 2000, Pazzani and Billsus, 1997] do not rely on ratings from other users, they can be used to produce recommendations for all items, provided attributes of the items are available. In fact, the content-based predictions of similar users can also be used to further improve predictions for the active user [Melville et al., 2002].

The new-user problem is difficult to tackle, since without previous preferences of a user it is not possible to find similar users or to build a content-based profile.

As such, research in this area has primarily focused on effectively selecting items to be rated by a user so as to rapidly improve recommendation performance with the least user feedback. In this setting, classical techniques from active learning can be leveraged to address the task of item selection [Jin and Si, 2004, Harpale and Yang, 2008].

Fraud

As Recommender Systems are being increasingly adopted by commercial websites, they have started to play a significant role in affecting the profitability of sellers. This has led to many unscrupulous vendors engaging in different forms of fraud to game recommender systems for their benefit. Typically, they attempt to inflate the perceived desirability of their own products (push attacks) or lower the ratings of their competitors (nuke attacks). These types of attack have been broadly studied as shilling attacks [Lam and Riedl, 2004] or profile injection attacks [Burke et al., 2005]. Such attacks usually involve setting up dummy profiles, and assume different amounts of knowledge about the system. For instance, the average attack [Lam and Riedl, 2004] assumes knowledge of the average rating for each item; and the attacker assigns values randomly distributed around this average, along with a high rating for the item being pushed. Studies have shown that such attacks can be quite detrimental to predicted ratings, though item-based Collaborative Filtering tends to be more robust to these attacks [Lam and Riedl, 2004]. Obviously, content-based methods, which only rely on a users past ratings, are unaffected by profile injection attacks.

While pure content-based methods avoid some of the pitfalls discussed above, Collaborative Filtering still has some key advantages over them. Firstly, CF can perform in domains where there is not much content associated with items, or where the content is difficult for a computer to analyze, such as ideas, opinions, etc. Secondly, a CF system has the ability to provide serendipitous recommendations, i.e. it can recommend items that are relevant to the user, but do not contain content from the user's profile.

3.2.3 Ranking in Recommendations

In many recommendation systems only the top-K items are shown to users. Recommendation is a ranking problem in the top-K recommendation situation. Ranking is more about predicting items' relative orders rather than predicting rating on items and it is broadly researched in information retrieval.

The problem of ranking documents for given queries is called Learning To Rank (LTR) [Liu, 2011] in information retrieval. If we treat users in recommendation systems as queries and items as documents, then we can use LTR algorithms to solve the recommendation problem. A key problem in using LTR models for recommendation is the lack of features. In information retrieval, explicit features are extracted from (query, document) pairs. Generally, three kinds of features can be used: query features, document features, and query-document-dependent features. In recommendation system, users' profiles and items' profiles are not easy to be represented as explicit features. Extracting efficient features for recommendation systems is emerged and also very important for learning a good ranking model.

Some work tries to extract features for user-item pairs [Volkovs and Zemel, 2012; Balakrishnan and Chopra, 2012] etc. In [Volkovs and Zemel, 2012], the authors extract features for a given *(u, i)* pair from user *u's* k-nearest neighbors who rated item *i*. They use rating-based user-user similarity metric to find neighbors for a target user. However, in some e-commerce systems users' preferences to items are perceived by tracking users' actions. A user can search and browse an item page, bookmark an item, put an item to the shopping cart and purchase an item. Different action indicates different preference to the item. For example, if a user *u* purchased item *i* and bookmarked item *j*, we can assume that user *u* prefers item *i* to item *j*. Mapping the user actions into a numerical scale is not natural and trivial. So it is hard to accurately compute the user-user similarity in e-commerce systems where users' feedbacks are non-numerical scores.

We will introduce ranking algorithms in recommendation system in this part. In [Cremonesi et al., 2010], by considering the missing entries as zeros, the authors perform conventional SVD (Singular Value Decomposition) on sparse matrix *R* and make top-K recommendation based on the value of the test items in the reconstruct matrix. Over-fitting on training dataset is avoided by only keeping high singular values. The algorithm shows good performance (precision/recall) than SVD++ [Koren, 2010], which is a famous model based on matrix factorization.

Some collaborative ranking algorithms are proposed based on LTR and matrix factorization. Ordrec [Koren and Sill, 2011] is an ordinal rank algorithm based on ordinal regression and SVD++. It models user's rating via SVD++ and aims at minimizing an ordinal regression loss. It brings the advantage to estimate the confidence level in each individual prediction and can also handle user's non-numerical feedbacks.

In [Rendle, 2009], the authors present a generic optimization criterion (BPR-OPT: Bayesian Personalized Ranking Optimization) for personalized ranking that is the maximum posterior estimator derived from a Bayesian analysis of the ranking problem. Bpropt optimizes the measure of the Area Under the ROC Curve (AUC) based on matrix factorization and adaptive KNN (*K*-Nearest Neighbors) algorithm. It uses stochastic gradient descent with bootstrap sampling to update parameters and converges very fast. PMF (Probabilistic Matrix Factorization) [Balakrishnan and Chopra, 2012] uses pairwise learning to rank algorithm to solve recommendation problem, both user-item features and the weights of the ranking function are optimized during learning. ListRankMF [Shi et al., 2010] aims at minimizing the cross entropy between the predict item permutation probability and true item permutation probability. In [Liu and Yang, 2008], the authors propose a probabilistic latent preference analysis model for ranking prediction by directly modeling user preferences by a mixture distribution based on Bradley-Terry model. Some collaborative ranking work tries to directly optimize evaluation measures. TFMAP [Shi et al., 2012] uses tensor factorization to model implicit feedback data with contextual information, and directly maximizes mean average precision under a given context.

In CLiMF (Collaborative Less-is-More Filtering) [Shi et al., 2012], the model parameters are learned by directly maximizing the Mean Reciprocal Rank and in CofiRank [Weimer et al, 2009], the authors fit a maximum margin matrix factorization model to minimize the upper bound of NDCG (Normalized Discounted Cumulative Gains). In [Volkovs and Zemel, 2013], the authors extract features for a given *(u, i)* pair from user *u*'s k-nearest neighbors who have rated item *i*. They use rating-based user-user similarity metric (e.g. Vector Space Similarity) to find neighbors for a target user. Their work has some drawbacks. First, they cannot well define user-user similarity when user's feedbacks are not numerical scores. Second, they choose neighbors by rating based user-user similarity metric, which is not corresponding to the ultimate goal of ranking.

3.2.4 Tagging and Social recommender systems

There has been a tremendous increase in user-generated content (UGC) in the past a few years via the technologies of Web 2.0. It is now well recognized that the user-generated content (e.g., product reviews, tags, forum discussions and blogs) contains valuable user opinions that can be exploited for many applications. By exploiting the UGC more effectively via the use of the latest collaborative filtering and data mining techniques, more accurate and sophisticate user profiles can be built which contain not only users' item preferences (i.e., item ratings) but also users' topic interests and trustworthiness between users. Based on the enhanced user profiles, high quality and reliable recommendations can be generated. Many significant researches have been done to investigate new strategies available in Web 2.0 framework. In this section, we review some new strategies for social recommender systems.

User generated content

Unlike the user rating data which is numeric data, the UGC comprises various forms of media and creative works such as written, audio, visual, and combined created by users explicitly and pro-actively. Therefore, it contains rich semantic information and provides a huge potential to obtain deeper knowledge about users, items, and the various relationships among users and items. It has become an important information resource in addition to traditional website materials. From the UGC information, it is possible to acquire users' opinions, perspectives, or tastes towards items or other users. The growing and readily available user-generated content is rising the new opportunity to construct user profiles accurately compared with the existing personalized recommender techniques and to mitigate the cold start and malicious rating problems considerably.

The UGC expresses users' opinions or sentiments towards items and is transforming how people seek advice and consider recommendations. The opinion mining and sentiment analysis such as customer opinion summarization [Zhuang et al., 2006] and sentiment analysis of user reviews [Ding et al., 2008] are possibly as augmentations to recommendation systems [Tatemura, 2000], since it might behoove such a system not to recommend items that receive a lot of negative feedback.

The individual users show their interest in online opinions about products or services. They share their brand experiences and opinions, positive or negative, regarding any product or service. The vendors of these items are increasingly coming to realize that these consumer voices can potentially wield enormous influence in shaping the opinions of other consumers and they are paying more and more attention to these issues [Hoffman, 2008]. There are already many companies that provide opinion mining services and examples include, Epinions.com, Amazon.com.

Blogs mining

As it is mentioned in chapter 2, the term web-log, or blog refers to a simple webpage consisting of brief paragraphs of opinion, information, personal diary entries, or links, called posts, arranged chronologically with the most recent first, in the style of an online journal. Most blogs also allow visitors to add a comment below a blog entry. People express their opinions, ideas, experiences, thoughts, and wishes through these free-form writings. A typical blog post can combine text, images, and links to other blogs, web pages, and other media related to its topic. The individuals who author the blog posts are referred as bloggers. The universe of all these blog sites is often referred as Blogosphere [Stewart et al., 2007]. Linking is also an important aspect of blogging as it deepens the conversational nature of the blogosphere and its sense of immediacy. It also helps to facilitate retrieval and referencing of information on different blogs.

Blogs notoriously contain quite a bit of subjective content. General topics include personal diaries, experiences, opinions, information technology, and politics to name a few. Thus blogs are more relevant than shopping sites for queries that concern politics, people, or other non-products. However, the desired material within blogs can vary quite widely in content, style, presentation. Mining opinions and sentiments from bloggers poses several challenges as compared to the historic feedback and surveys.

State-of-the-art content analysis techniques could be used for basic clustering, classification of the blog posts/blog sites. For example, a prototype system called Pulse [Gamon et al., 2005] uses a Naive Bayes classifier trained on manually annotated sentences with positive/negative sentiments and iterates until all unlabeled data is adequately classified.

The researchers [Joshi and Belsare, 2006] developed a blog mining program called Blog-Harvest which searches for, and extracts, a blogger’s interests in order to recommend blogs with similar topics. The program uses classification, links, topic similarity clustering and tagging based on opinion mining to provide these features. The program design is based on the knowledge that blogging communities are not formed randomly, but as a result of shared interests. It is also designed to provide a useful search facility to bloggers while generating large amounts of revenue for advertising services and providers.

Tag-based recommender systems

Recommender systems in general recommend interesting or personalized information objects to users based on explicit or implicit ratings. Usually, recommender systems predict ratings of objects or suggest a list of new objects that the user hopefully will like the most.

The approaches of profiling users with user-item rating matrix and keywords vectors are widely used in recommender systems. However, these approaches are used for describing two-dimensional relationships between users and items. In tag recommender systems the recommendations are, for a given user u ∈ U and a given resource r ∈ R, a set T' (u, r) ⊆ T of tags. In many cases, T' (u, r) is computed by first generating a ranking on the set of tags according to some quality or relevance criterion, from which then the top *n* elements are selected [Jäschke et al., 2007].

Personalized recommendation is used to conquer the information overload problem, and collaborative filtering recommendation is one of the most successful recommendation techniques to date. However, collaborative filtering recommendation becomes less effective when users have multiple interests, because users have similar taste in one aspect may behave quite different in other aspects. Information got from social tagging websites not only tells what a user likes, but also why he or she likes it.

In the remainder of this section, we first describe the extension with integrating tags information to improve recommendation quality. We then present well-known recommendation algorithms for developing folksonomies.

a) Extension with tags

The current recommender systems are commonly using collaborative filtering techniques, which traditionally exploit only pairs of two-dimensional data. As collaborative tagging is getting more widely used social tags as a powerful mechanism that reveals three-dimensional correlations between users–tags–items, could also be employed as background knowledge in Recommender System.

The first adaptation lies in reducing the three-dimensional folksonomy to three two-dimensional contexts: <user, tag> and <item, tag> and <user, item>. This can be done by augmenting the standard user-item matrix horizontally and vertically with user and item tags correspondingly [Tso-Sutter et al., 2008]. User tags, are tags that user *u* uses to tag items and are viewed as items in the user-item matrix. Item tags, are tags that describe an item *i* by users and play the role of users in the user-item matrix. Furthermore, instead of viewing each single tag as user or item, clustering methods can be applied to the tags such that similar tags are grouped together.

Supporting users during the tagging process is an important step towards easy-to-use applications. Consequently, different approaches have been studied in the past to find best tag recommendations for resources.

b) Recommender systems and folksonomies

Despite the relative newness of folksonomies, there are a lot of recommender approaches attached to this domain. Most of these contributions are distributed between tag recommendation and resource recommendation. In the following subsections, we will give an overview about the main contributions related to our work.

Tag Recommendation: The general aim of tag recommender systems is to help users choose the appropriate tags when annotating resources. Among the many works addressing this problem, let us cite that of [Schmitz et al., 2006] who showed how association rules can be adopted to analyze and structure folksonomies and how these folksonomies can be used for learning ontologies and supporting emergent semantics. Another noticeable contribution is that of [Jäschke et al., 2007] who presented a formal model and a new search algorithm called FolkRank, especially designed for folksonomies. It is also applied to find communities within a folksonomy and is used to structure search results. [Gemmell et al., 2009] proposed a tag-based recommendation method based on the adaptation of the K-nearest neighbors algorithm so that it accepts as input both a user and a resource and gives out a set of tags. The interest of this approach is to orient users to use the same tags, and thus increase the chance of building a common vocabulary used by all the community members.

Resource Recommendation: The general aim of resource recommender systems is to insure the quantity and relevance of the recommended resources. Among the works addressing this problem, let us cite [Tso-Sutter et al., 2008], who described a method that allows tags to be incorporated into standard heuristic-based collaborative filtering algorithms, and apply a fusion method to re-associate these correlations to recommend resources. [Zhao et al., 2008] proposed a Clustered Social Ranking (CSR), a new search and recommendation technique specifically developed to support new users of social websites finding contents. The system detects who the leaders are; it then clusters them into communities. User queries are then directed to the community of leaders who can best answer them.

[De Meo et al., 2010] proposed an approach based on the principle of query expansion to enrich user profiles by additional tags discovered through the exploration of the two graphs: Tag Resource Graph (TRG) and Tag User Graph (TUG) representing the relations respectively between tags and resources and between tags and users.

[Huang et al., 2011] proposed a recommender system that considers the user recent tag preferences. The proposed system includes the following stages: grouping similar users into clusters, finding similar resources based on the user resources, and recommending the top-N items to the target user.

[Zanardi and Capra, 2011] proposed a method aimed to extend the searching capabilities of digital collections targeting educational and academic domains. Given a document, the approach finds similar documents that may be relevant to the user. [Versin et al., 2013] developed a personalized web-based recommender system that applies recommendation and adaptive hypermedia techniques to orient learner's activities and recommend pertinent links and actions to him during learning. The proposed approach is based on using data clustering, collaborative filtering and association rule mining techniques. A summary of those contributions is presented in table 3.1.

In chapter 4, we aim propose a new approach to recommend personalized resources base on association rules. Personalization can significantly improve folksonomy-based recommender systems, because the man-machine interaction and therefore the user effort are considerably reduced.

3.2.5 Group Recommender Systems for the Social Web

Social systems encourage interaction between users and both online content and other users, thus generating new sources of knowledge for recommender systems. The Social Web presents thus new challenges for recommender systems [Geyer et al., 2010]. In the context of group recommendations, we can highlight the following research directions:

-Developing new applications. The huge amount and diversity of user generated content available in the social web allow investigating scenarios in which a group of individuals is recommended with “social objects” such as photos, music tracks and video clips stored in online multimedia sharing sites; stories, opinions and reviews published in blogs; and like-minded people registered in online social networks. In such applications, user generated content like ratings, tags, posts, personal bookmarks and social contacts could be exploited by novel group recommendation algorithms [Geyer et al., 2010].

-Dealing with dynamics and diversity of virtual communities. In online social networks, people tend to reproduce or extend their relations in the real world to the virtual worlds conformed by the social networks. In [Szomszor et al., 2010], the authors show that relationship strength can be accurately inferred from models based on profile similarity and interaction activity on online social networks. Based on these findings, group recommender systems could incorporate content and social interests of group members to perform more accurate item suggestions. For such purpose, it would be necessary to investigate large group characteristics that impact individual decisions, and explore new satisfaction and consensus functions that capture social, interest, and expertise (dis)similarity among the members of a community [Gartrell et al., 2010]. With this respect, because of the evolving composition of online communities, analyzing and exploiting the time dimension in the above characteristics may play a key role to obtain more accurate recommendations for community members.

-Incorporating contextual information. The anytime-anywhere phenomenon is present in any social system and thus, group recommenders for the social web should incorporate contextual information [Adomavicius and Tuzhilin, 2011]. They would have to automatically detect user presence from inputs provided by mobile, sensor and social data sources [Szomszor et al., 2010], and adaptively infer the strength of the social connections within the group, in order to provide accurate recommendations.

-Finding communities of interest. In the social web, it is very often the case that the membership to a community is unknown or unconscious. In many social applications, a person describes her interests and knowledge in a personal profile to find people with similar ones, but she is not aware of the existence of other (directly or indirectly) related interests and knowledge that may be useful to find those people. Furthermore, depending on the context of application, a user can be interested in different topics and groups of people. In both cases, for individual and group recommender systems, a strategy to automatically identify communities of interest could be very beneficial [Cantador and Castells, 2011].

-Integrating user profiles from multiple social systems. Increasingly, users maintain personal profiles in more and more Web 2.0 systems, such as social networking, personal bookmarking, collaborative tagging, and multimedia sharing sites. Recent studies have shown that inter-linked distributed user preferences expressed in several systems not only tend to overlap, but also enrich individual profiles [Szomszor et al., 2008; Winoto and Ya Tang, 2008]. A challenging problem in the recommender system field is the issue of integrating such sources of user preference information in order to provide the so called cross-domain recommendations [Winoto and Ya Tang, 2008]. This clearly opens new research opportunities for group recommenders, which e.g. could suggest to a virtual community sharing interests in a particular domain with items belonging to other domain but liked by some of its members.

3.2.6 Medical recommendation system in Social Tagging Activities

Health social networking is a newly emerging health care service that facilitates information exchange and collaboration within the communities of patients and caregivers. One of the challenges of health social networking is finding and recommending communities that patients can share their stories with and obtain social and emotional support from, as well as, clinical knowledge to enhance their self-care. In this section we report the recent approaches discussed this issue.

The approach of [Lim and Husain, 2010] conveys web-based wellness system that enables users to search for personalized wellness therapy for them. The authors suggest an implementation of wellness community portal that supported by a case-based reasoning (CBR) wellness recommender system. CBR wellness recommender system is to be used by users in searching for personalized wellness therapy in the Internet.

In the contribution of [Lopez-Nores et al., 2011], the authors have enhanced an existing recommender system to process information stored in electronic health records and in groups of interest created within social networks.

The solution proposed in [Donciu et al., 2011] is a recommender system for running professionals and amateurs, which is able to provide information to users regarding the workout and the diet that best suits them, based on their profile information, preferences and declared purpose. The solution mixes a social dimension derived from an expanding community with expert knowledge defined within an ontology. Moreover, the proposed model addresses adaptability in terms of personal profile, professional results and unfortunate events that might occur during workouts.

In [Song and Marsh, 2012], the authors developed a new information-sharing scheme where each patient is represented as a small number of (possibly disjoint) d-words (discriminantwords) and the d-words are used to measure similarities between patients without revealing sensitive personal information. The d-words are simple words like “food,” and thus do not contain identifiable personal information. The d-words can be easily shared on the Internet to find peers who might have similar health conditions.

In [Hamed et al., 2012], the authors introduced their initial work for developing a social networks recommender system called T-Recs. The system is a time-aware Twitter-based alternative medicine recommender system. They collected a set of tweets that contain specific hashtags, and then the individual tweets were examined manually by a domain expert (a medical doctor) who inspected the tweet sentiments along with the tweet’s timestamp. Using this data, the domain expert assigned a preliminary label to the tweet/group of tweets. Next, the authors trained a classifier using the hashtags and its labels taking into consideration other factors (i.e., age, gender, co-morbidity conditions) that the Tweeter must provide by taking a questionnaire. After the questions are answered, the recommender system makes recommendations based on the Tweet contents and the questionnaire factors for never-seen-before tweets. The classifier, is the core component of the recommender system, is designed as a Decision Tree algorithm to classify the Tweets. Each recommendation is made by the system provides a medical advice to promote public health awareness, and a link to the recommender systems' web portal to answer the questionnaire. When the questionnaire is submitted, the symptoms and Tweeter's basic information are consolidated and a recommendation is made available to the user at once.

The work of [Zaini et al., 2012] presents an online system that promotes audio therapy for tuning users' mental states to different modes. By using this system, users may tune and listen to meditative audios to get the calming effects to relieve stress, listening to energizing audios when feeling lethargic in the morning and listening to enlightening audios to help in learning. Being a first timer to use such system, a new user may need assistance in choosing a suitable and effective audio in getting the desired outcome. According the authors, one way in offering such assistance is to deploy a recommendation platform, where ratings from other users of the system; (especially friends of the user) are counted and adopted as the basis for recommendations. The novelty of this system lies in the mechanism for constructing and personalizing the recommendations on suitable therapy audios for a particular user.

In the work of [Rivero-Rodríguez et al., 2013], the authors proposed a Health Information Recommender System to connect videos with trustworthy information from very trustful medical sources, such as Medline Plus. According to video's data, this system detects the main topic of the video and enriches it with information from very well-known resources.

In their contribution, [Zaman and Li, 2014] proposed a recommendation system which utilizes semantic web technology and healthcare social networking to provide personalized recommendation to speed patient recovery and improve healthcare outcomes.

[Li and Zaman, 2014] proposed a personalized healthcare recommending system to recommend highly relevant and trustworthy healthcare –related information to users. The system identifies key factors impacting the recommendation in a healthcare social networking environment, and uses semantic web technology and fuzzy logic to represent and evaluate the recommendation.

The growing amount of user generated content in healthcare social media requires new methods to gain new insights about patients as users of online health communities.

One interesting general problem is filtering information and making relevant information more conspicuous to users in Internet websites. Previously, collaborative filtering techniques have been successfully applied in recommender systems for personalizing Internet shopping and movie portals. Current work uses network structures – inferred from patient interactions– to enhance patient-similarity analysis, for predicting the top-N threads in online communities. Using network structure properties unique to healthcare social networks, [Chomutare, 2014] performed experimental results based on the Euclidean distance, Pearson correlation and Tanimoto similarity. These findings have implications for designing personalized health-related social media and confirm that community structure properties can enhance recommendations.

In [Zhang and Yang, 2014], the authors utilized Naïve Bayes methods and proposed two tasks for classifying posts and comments on QuitStop forum, an online community for smoking cessation intervention, respectively: (1) classification of intentions and (2) classification of social support types. Different text feature sets and user health feature sets are selected to develop classifiers. Taking different evaluation indicators as optimizing goals, the authors developed genetic algorithms to combine classifiers with different feature sets and optimize the classification results. For comment classification, combining different text features could reach the best result. In the future, the classification result could be applied to developing recommender systems for topic recommendation and user prediction of online health forums.

The health care support system is a special type of recommender systems that play an important role in medical sciences nowadays. This kind of systems often provides the medical diagnosis function based on the historic clinical symptoms of patients to give a list of possible diseases accompanied with the membership values. The most acquiring disease from that list is then determined by clinicians' experience expressed through a specific defuzzification method. An important issue in the health care support system is increasing the accuracy of the medical diagnosis function that involves the cooperation of fuzzy systems and recommender systems in the sense that uncertain behaviors of symptoms and the clinicians' experience are represented by fuzzy memberships whilst the determination of the possible diseases is conducted by the prediction capability of recommender systems. Intuitionistic fuzzy recommender systems (IFRS) are such the combination, which results in better accuracy of prediction than the relevant methods constructed on either the traditional fuzzy sets or recommender system only. Based upon the observation that the calculation of similarity in Intuitionistic fuzzy recommender systems (IFRS) could be enhanced by the integration with the information of possibility of patients belonging to clusters specified by a fuzzy clustering method, in the work of [Thong and Son, 2015], the authors proposed a novel hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis so-called HIFCF (Hybrid Intuitionistic Fuzzy Collaborative Filtering). The significance and impact of this new method contribute not only the theoretical aspects of recommender systems but also the applicable roles to the health care support systems.

In the work of [Gong et al., 2014], the authors tried to systematically study how to perform doctor recommendation in mobile Medical Social Networks (*m*-MSNs). Specifically, employing a real-world medical dataset as the source in their study, they first mine doctor-patient ties/relationships via Time-constraint Probability Factor Graph model (TPFG), and then define the transition probability matrix between neighbor nodes. Finally, the authors propose a doctor recommendation model via Random Walk with Restart (*RWR*), namely *RWR*-Model.

In [Song et al., 2011], the authors proposed a social-networking framework for patient care, in particular for parents of children with Autism Spectrum Disorders (ASD). In the framework, health service providers facilitate social links between parents using similarities of assessment reports without revealing sensitive information. A machine learning approach was developed to generate explanations of ASD assessments in order to assist clinicians in their assessment. The generated explanations are then used to measure similarities between assessments in order to recommend a community of related parents.

It is difficult for patients to find the most appropriate doctor/physician to diagnose. In most cases, just considering Authority Degrees of Candidate Doctors (AD-CDs) cannot satisfy this need due to some objective preferences such as economic affordability of a patient, commuting distance for visiting doctors and so on. In their work, [Gong and Sun, 2011] tried to systematically investigate the problem and propose a novel method to enable patients access such intelligent medical service like this. In the method, the authors first mine patient-doctor relationships via Time-constraint Probability Factor Graph mode (TPFG) from a medical social network, and then extract four essential features for AD-CDs that would be subsequently sorted via Ranking SVM. At last, combining AD-CDs and patients' preferences together, the authors propose a novel Individual Doctor Recommendation Model, namely IDR-Model, to compute doctor recommendation success rate based on weighted average method.

In chapter 5, a novel technique will be proposed to help doctors discovering the best practices to patient diseases diagnosis and treatments in their daily tasks by analyzing doctors' profiles according to their tagging activity in order to personalize a greatest medical-resources recommendation related to the patients' diseases, treatments or clinical cases. We propose to take profit of community effect strength which characterizes social networks with creating and observing emergence of the intelligence captured from socials interactions between doctors in the network by using a powerful method of Data Mining which is Associations Rules. We show through an empirical scenario how we can evaluate and demonstrate the efficiency of the medical resource recommender system in clinical decision.

3.2.7 Areas for further research

As described above, there has been much research done on tag-based recommendation technologies, over the past several years that have used a broad range of information retrieval, and other techniques that have significantly advanced the state-of-the-art in comparison to early recommender systems that utilized collaborative and content-based heuristics.

There are also clearly great benefits in user tagging and folksonomies, especially in the richness, currency, relevance and diversity of the terms used, and the collections of resources created. The success of tagging services like Flickr, Delicious, etc. has shown that tagging is a great collaboration tool. Tagging seems to be the natural way for people to classify objects as well as an attractive way to discover new material. Tagging services provides users with a repository of tagged resources that can be searched and explored in different ways. More and more people use at least one tagging service and enjoy them as discovery tools.

The rapid development of collaborative tagging system and related emerging technology suggests new ideas for personalized recommendation and determine a great number of challenges for future work. Interesting additional features [Firan et al. 2007; Xu et al. 2006], which are worth for further research, are listed below.

Incorporate relevance feedback into search-based recommendations, such that the user is able to select negative tags or items he does not like.

Improve tag browsing experience by applying the same principles in constructing tag cloud, e.g., by presenting tags with good facet mix while considering popularity and user interests.

Examining the distribution of tag use, that is, the most used tags are more likely for other users since they are more likely to be seen. There will be a few tags that are used by a substantial number of users. An order of magnitude more tags that are used by fewer users, and another order of magnitude more used by only a handful of users. Examining this sort of distribution of tag use could give a better indication of whether a folksonomy converges on terms, or the distribution of terms flattens, perhaps indicating less agreement.

Examining user behavior through ethnographic observation or interview to understand his motivations and cognitive processes in tagging items. Although it seems that some users are intending to facilitate communication through tag use, especially in the unintended uses, interviews could make this point explicit. Interviews could also elucidate the conscious intentions of users in “normal” use of the system, which is much harder to observe simply from the documents and tags themselves.

Analysis of the frequency which users modify or change their tags, or future tagging behavior based on the implicit feedback from the system in the form of what other documents are tagged with a term.

The use of a folksonomy to supplement existing classification schemes and provide additional access to materials by encouraging and leveraging explicit user metadata. The organizational schemes developed by the users have the possibility to be of great interest to other users and improve the recommendation systems.

Improve tag uniformity by normalizing semantically similar tags that are not similar in letters.

These extensions leave ample opportunity for future work in this area. They can improve tag-based recommendation capabilities and make collaborative tagging systems applicable to en even broader range of applications. Thus, in this thesis we propose a new approach to improve resource recommendation within folksonomies.

3.3 Conclusion

In the first section of this chapter, we presented an exhaustive description about social web search; this effective approach directly retrieves documents when querying search engines. The presented details including collaborative tagging, personalization and ranking demonstrated that tagging activities has a positive impact on social information retrieval.

Recommender systems made significant progress over the last decade when numerous content-based, collaborative, and hybrid methods were proposed and several systems have been developed. However, despite all of these advances, the current generation of recommender systems still requires further improvements to make recommendation methods more effective in a broader range of applications. With the increasing popularity of the collaborative tagging systems, tags could be interesting and useful information to enhance recommender systems’ algorithms. Besides helping user organize his or her personal collections, a tag also can be regarded as a user’s personal opinion expression, while tagging can be considered as implicit rating or voting on the tagged information resources or items. Thus, the tagging information can be used to make recommendations. In this chapter, we described social tagging systems which can be used for extending the capabilities of recommender systems. We presented a comprehensive survey of the state-of-the-art in collaborative tagging systems, folksonomy and tag-based recommender systems. Various limitations of the current generation of folksonomy systems and possible extensions that can provide better recommendation capabilities are also considered.

To sum up, in this chapter we described the two information access paradigms that users undertake each time they need to meet particular information needs on the web: searching by query and recommendation. The state of the art of these paradigms demonstrates the existence of some gaps. In the following chapters we will try to overcome these insufficient.

	Social web														
	Tag recommendation			Resource recommendation						Social Semantic web Approaches					
	[Schmitz et al., 2006]	[Jaschke et al., 2007]	[Gemmell et al., 2009]	[Tso-Sutter et al., 2008]	[Zhao et al., 2008]	[De Meo et al., 2010]	[Huang et al., 2011]	Zanardi and Corpa, 2011]	[Versin et al., 2013]	[Mika, 2005]	[Gruber, 2005]	[Specia and Motta, 2007]	[Limpens et al., 2010]	[Pan et al., 2010]	[Wu and Zhou, 2011]
Tags Ambiguity	-	-	-	-	-	-	-	-	-	-	+	-	-	+	+
Spelling variations	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+
Folksonomy Enrichment	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
Personalization	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-
Social similarities	-	-	-	+	+	+	+	+	+	+	-	-	-	-	+
Tags similarities	+	+	+	+	-	+	+	+	+	+	+	+	+	-	+
Scalability	-	-	+	+	+	+	+	+	+	+	-	+	+	-	+
Users Contributions	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
Automatic	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
Semantic Annotations	-	-	-	-	-	-	-	-	-	+	+	+	+	+	-

Table 3.1: Summary of some functionalities in many social web approaches. (-) means the dimension (i.e. functionality) is provided, (-) means the dimension is not provided. These marks do not imply any “positive” or “negative” information about the tools except the presence or the absence of the considered dimension

Part 2:
Contributions

Chapter 4:
Personalizing and Improving Resource Recommendation in Folksonomies

Chapter 4:

Personalizing and Improving Resource Recommendation within Folksonomies

Collaborative tagging which is the keystone of the social practices of web 2.0 has been highly developed in the last few years. In this chapter, we propose a new method to analyzing user profiles according to their tagging activity in order to improve resource recommendation. We base upon association rules which are a powerful method to discover interesting relationships among large datasets on the web. Focusing on association rules we can find correlations between tags in a social network. Our aim is to recommend resources annotated with tags suggested by association rules, in order to enrich user profiles. The effectiveness of the recommendation depends on the resolution of social tagging drawbacks. In our recommender process, we demonstrate how we can reduce tag ambiguity and spelling variations problems by taking into account social similarities calculated on folksonomies, in order to personalize resource recommendation. We surmount also the lack of semantic links between tags during the recommendation process.

In another side, when inquiring folksonomies a large resources number can match users' queries. Therefore, searching and ranking these web resources is required to release user who cannot browse all them. Generally, search results are displayed in chronological order. This ordering is not always effective; therefore a retrieval model should rank resources by the degree of relevance with respect to time. An event-aware ranking model takes into account event detection, which captures the fact that the relevance of resources may change over time according to new events detection. We want to study whether exploiting other features together with event detection can help improving the retrieval effectiveness in searching resources within social applications. In this case, we need to find features used for capturing the similarity between an information need and personalization of retrieved resources, and combine such features with event dimension for relevance ranking. In this context, a social personalized ranking function is proposed to leverage the social aspect of folksonomy and events detection. The aim is to acquire relevant resources recommended to or inquired by users’ tag-based query.

4.1 Introduction

With web 2.0 technologies the web has become a social space where users create, annotate, share and make public resources which they find interesting on the web [Beldjoudi et al., 2011b]. Andreas Kaplan and Michael Haenlein define social media as "a group of Internet-based applications that build on the ideological and technological foundations of web 2.0, and that allow the creation and exchange of user-generated content" [Kaplan and Haenlein, 2010]. Folksonomies are one of the keystones of these new social practices: they are systems of classification resulting from collaboratively creating and managing tags to annotate and categorize contents. This practice is known as collaborative tagging or social tagging. The basic principle of social tagging relies on three main notions: the user, the resource and the tag. The combination of these three elements enables exploiting annotations of web resources by users with tags.

Despite the strength of folksonomies, there are some problems hindering the growth of these systems: tag ambiguity (or polysemy) is one of the famous problems in folksonomies. It comes from the fact that a tag can designate several concepts (i.e., a tag can have several meanings), for example when a user employs the tag "apple" to annotate a resource, the system will not understand if the user means the fruit or the company. Also the variations in writing a same concept (spelling variations or synonymy) can cause some problems during the search phase, for example "cat" and "chat" both denote the same concept (animal) in English and in French, but when a user searches resources annotated by the tag "cat", the system will not offer him those tagged with the word "chat" because it cannot understand that the tag "cat" has the same meaning that the tag "chat". In addition, tags that are freely chosen in these systems are likely to contain spelling errors and therefore make the retrieval of resources more doubtful than the metadata recovering from a lexicon examined by information professionals. Therefore resource retrieval within folksonomies needs some improvements to increase the quality of the results obtained in these systems.

In this chapter, we propose a method to analyze user profiles according to their tags in order to predict interesting personalized resources and recommend them. In other words, our objective is to enrich the profiles of folksonomy users with pertinent resources. We argue that the automatic sharing of resources strengthens social links among actors and we exploit this idea to reduce tag ambiguity and spelling variations in the recommendation process by increasing the weights associated to web resources according to social similarities. We base upon association rules which are a powerful method for discovering interesting relationships among a large dataset on the web. We insist on the fact that our final aim is not to suggest tags to users: each time a resource is presented to a user, the tags already used to annotate this resource are indicated but the user is free to tag the resource by choosing a tag among them or by using a new one. Our aim is to recommend resources which are annotated with tags suggested by association rules, in order to enrich user profiles with these resources. Our approach comes from a new view on the community effect in folksonomies since it aims at automatically strengthening existing correlations between different members of online communities, without involving the user in this process. The fact of suggesting to each user some resources considered useful or interesting for him without specifying explicit tags, this can significantly improve folksonomy-based recommender systems, because the man-machine interaction and therefore the user effort are considerably reduced.

In folksonomies a relatively large number of resources can match users' queries. Therefore, ranking these web resources is a key problem, since a user cannot browse all these resources. Ranking in information retrieval consists of the definition of a ranking function that allows quantifying the similarities among documents and queries [Bouadjenek et al., 2013].

To define a valued ranking function in collaborative environment we focused on two features: social aspect and event detection. In the first feature, we base our approach on similarities calculation to overcome tags ambiguity problem. While in the second feature, we introduce a new dimension which is event detection to define the impact of resources popularity on ranking the retrieved resources. Consequently, we want to benefit from the dependence between the presence of an event and the high number of similar queries transmitted in the same period by different users in order to improve resources retrieval in collaborative applications.

The rest of this chapter is organized as follows: Section 4.2 is dedicated to our approach concerning resources' recommendation in folksonomies. In Section 4.3 we will present a personalized ranking function in collaborative environment. Conclusions are described in Section 4.4.

4.2 Resource Recommendation in Social Networks

Our objective is to develop a new approach based on social interactions between different members in a community to make semantics emerge in folksonomies, with the aim of personalizing resource recommendation. The key idea of our approach is to make each member benefit from the resources tagged by other users who have similar interests. We measure similarity between community members in order to compare their preferences and then suggest relevant resources. This allows limiting the problems of ambiguity, spelling variations and the lack of semantic links between tags in folksonomies.

Our approach comes with a new view on the community effect in folksonomies, which consolidates the social interactions between the different members of a community without involving the user in the automatic realization of this process. Also, the fact of proposing to each user resources considered useful to him without him identifying specific tags can significantly improve folksonomy-based systems, because this reduces the man-machine interactions. The user effort is reduced to a mouse click instead of a keyboard input and therefore this should encourage users to use these systems.

4.2.1 Approach Description

We define a folksonomy by a tripartite model where web resources are associated with a user to a list of tags. Formally a folksonomy is a tuple F = <U, T, R, A> where U, T and R represent respectively a set of users, a set of tags and a set of resources, and A represents the relationships between the three preceding elements, i.e. $A \subseteq U \times T \times R$ [Mika, 2005].

We extract three social networks from a folksonomy, which represent three different viewpoints on social interactions: one network relating tags and users, a second one relating tags and resources and a third one relating users and resources. We represent these social networks by three matrices TU, TR, UR:

-TU = [X_{ij}] where : $X_{ij}= \begin{Bmatrix} 1 \text{ if } \exists\ r \in R, < uj, ti, r > \in A \\ 0 \text{ otherwise} \end{Bmatrix}$

-TR = [Y_{ij}] where: $Y_{ij} = \begin{Bmatrix} 1 \text{ if } \exists\ u \in U, < u, ti, rj > \in A \\ 0 \text{ otherwise} \end{Bmatrix}$

-UR = [Z_{ij}] where: $Z_{ij} = \begin{Bmatrix} 1 \text{ if } \exists\ t \in T, < ui, t, rj > \in A \\ 0 \text{ otherwise} \end{Bmatrix}$

RU, RT and UT are the transposed matrices of UR, TR and TU.

This enables us to analyze the correlations captured from the different social interactions. We use Pajek, a tool which has already been used by Mika to analyze large networks [Mika, 2005].

To apply an association rule method to folksonomies, we represent each user in a folksonomy by a transaction ID and the tags he uses by the set of items which are in this transaction [Beldjoudi et al., 2011b]. Table 4.1 provides an illustrative example of a dataset of user tags.

Transaction ID	Itemset
U_1	Computer, Programming
U_2	Computer, Apple
U_3	Kitchen, Apple
U_4	Programming
U_5	Kitchen

Table 4.1: An illustrative example of a dataset with user tags

Our goal is to find correlations between tags, i.e. to find tags frequently appearing together, in order to extract those which are not used by one particular user but which are often used by other users close to him in the social network.

For example, let us consider a dataset in which it occurs that many users who use the tag Software also employ the tag Java. We aim at extracting a rule Software ⇒ Java so that we can enrich the profiles of users who employ the tag Software but not the tag Java, by the resources tagged with Java. Among the wide range of algorithms proposed to extract interesting association rules, we use the one known as Apriori [Agraval et al., 1993].

Once the rules are extracted, our recommender system proceeds as follows: For each extracted rule, we test whether the tags which are in the antecedent of the rule are used by the current user. If it is the case then the resources tagged with each tag found in the consequent of the rule are candidate to be recommended by the system. The effectiveness of the recommendation depends on the resolution of folksonomies problems. In our approach we tackle the problems of tag ambiguity, spelling variations (or synonymy) and the lack of semantic links between tags. The detail of our approach is described in the following subsections.

4.2.2 Resolving Tags Ambiguity in Recommendation

According to [Mathes, 2004], "The problems inherent in an uncontrolled vocabulary lead to a number of limitations and weaknesses in folksonomies. Ambiguity of the tags can emerge as users apply the same tag in different ways. At the opposite end of the spectrum, the lack of synonym control can lead to different tags being used for the same concept, precluding collocation".

A tag can have several meanings, i.e. refer to several concepts. Therefore, a basic tag-based recommender system would equally recommend resources relative to fruits or to computers for a user searching with the tag "apple". The resolution of tag ambiguity is especially crucial in our approach where some tags which are used to recommend resources are not directly used by the user but deduced with association rules. To resolve the problem of tag ambiguity in recommendation, we propose to measure the similarity between users to identify those who have similar preferences and therefore adapt the recommendation to user profiles [Beldjoudi et al., 2011b].

First step: For each extracted association rule A ⇒B whose antecedent applies to an active user u_x, we measure the similarities between this user and the users of his social network who use the tags occurring in the consequent of the rule (see figure 4.1). The resources associated to these tags are recommended to the user depending on these similarities. To measure similarity between two users u1 and u2, both are represented by a binary vector representing all their tags (extracted from matrix UT: see figure 4.1) and we compute the angle cosines between the two vectors:

$$\text{sim}(u_1, u_2) = cos(v_1, v_2) = \frac{v_1.v_2}{\|v_1\|^2\|v_2\|^2} \qquad (4.1)$$

According to [Cattuto et al., 2008] and [Koerner et al., 2010], the cosine similarity gives good quality results at a reasonable computational cost since it has linear complexity.

We insist on the fact that the distribution of tags over resources and users in folksonomies follows a power law [De Meo et al., 2010]: most resources are tagged by only a small number of users, and many tags are only used by a few users, a property which leads to a low values of r (the number of resources in matrix RU: see figure 4.1) and n (the number of users in the matrix UT: see figure 4.1). Therefore, our approach can scale to very large datasets.

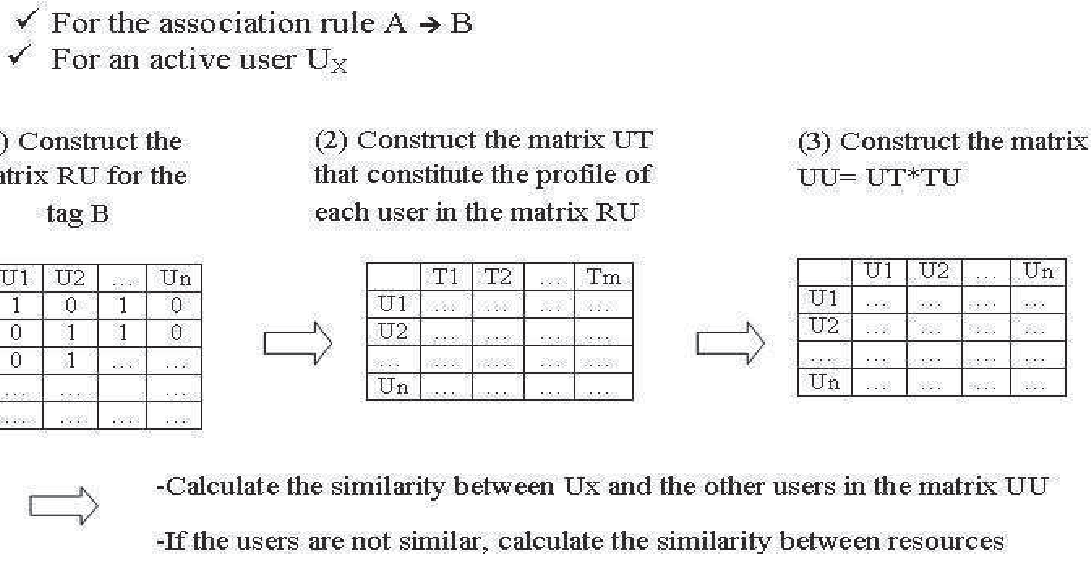
Second step: To avoid the cold start problem which generally results from a lack of data required by the system in order to make a good recommendation, when the user of the recommender system is not yet similar to other users, we also measure the similarity between the resources which would be recommended by the system (as related to a tag occurring in the consequent of an association rule) and those which are already recommended to the user. To measure the similarity between two resources r1 and r2, we represent each of them by the binary vector representing all its tags (extracted from matrix TR) and we calculate the cosines of the angle between the two vectors.

Third step: Each resource recommended by the system is first associated an initial weight based on the similarities between users. Above a threshold fixed in [0..1], we qualify the resource as highly recommended. Under this threshold, we consider the similarity between resources and we similarly highly recommend the resources which weights calculated on the product matrix RR = RT x TR are above a given threshold.

We note that our recommender system is flexible, since the user can interact to accept or reject the recommended resources. Also, the very power low distribution of resources over users in folksonomies leads to a low value of *r* (the number of resources in matrix RU). Therefore, the product matrix RR = RU x UR is not expensive in our case, which makes the approach efficient and scale to very large datasets.

For instance, let us consider a folksonomy with five users who annotate five resources using four tags. Each triple (u, t, r) represents a connection between a user, a resource and a tag (see table 4.2).

Let the extracted association rule computer ⇒ apple and the folksonomy described in table 4.2. Since the tag "computer" is used by user U1, then resources R3 and R5 tagged with the tag "apple" (in the consequence of the rule) are candidates for a recommendation to U1. Matrix UT (table 4.4) shows that "apple" is used by users U2 and U3. Then we calculate the similarity between U1 and U2 and the similarity between U1 and U3, based on matrix UU = UT x TU (table 4.5).



- ✓ For the association rule A → B
- ✓ For an active user U_X

(1) Construct the matrix RU for the tag B

	U1	U2	…	Un
R1	1	0	1	0
R2	0	1	1	0
R3	0	1	…	…
…	…	…		…
Rr	…	…	…	…

(2) Construct the matrix UT that constitute the profile of each user in the matrix RU

	T1	T2	…	Tm
U1	…	…	…	…
U2	…	…	…	…
…	…	…	…	…
Un	…	…	…	…

(3) Construct the matrix UU= UT*TU

	U1	U2	…	Un
U1	…	…	…	…
U2	…	…	…	…
…	…	…	…	…
Un	…	…	…	…

-Calculate the similarity between Ux and the other users in the matrix UU

-If the users are not similar, calculate the similarity between resources

Figure 4.1: Process of surmounting the tag ambiguity problem

Users	Tags	Resources
U1	computer	R1
U1	programming	R2
U2	computer	R1
U2	Apple	R3
U3	Kitchen	R4
U3	Apple	R5
U4	programming	R1
U5	Kitchen	R4
U5	Kitchen	R5

Table 4.2: Example of a folksonomy

	U2	U3
R3	1	0
R5	0	1

Table 4.3: Matrix RU of tag apple

	Computer	Kitchen	programming	apple
U1	1	0	1	0
U2	1	0	0	1
U3	0	1	0	1

Table 4.4: Matrix UT

	U1	U2	U3
U1	2	1	0
U2	1	2	1
U3	0	1	2

Table 4.5: Matrix UU

$sim(U_1,U_2) = cos(UU_1,UU_2) = \frac{(2\ 1\ 0) \text{ x } (1\ 2\ 1)}{\sqrt{4+1+0} \text{ x } \sqrt{1+4+1}} = \frac{4}{\sqrt{30}} = 0.73$

$sim(U_1,U_3) = cos(UU_1,UU_3) = \frac{(2\ 1\ 0) \text{ x } (0\ 1\ 2)}{\sqrt{4+1+0} \text{x } \sqrt{0+1+4}} = \frac{1}{5} = 0.2$

U1 and U2 show higher cosine similarity than U1 and U3. Then, among the resources tagged with "apple", namely R3 and R5, those tagged by U2 are highly recommended to U1: it is the case of R3.

U1 and U3 are not similar, then, among the resources tagged with "apple", we compute the similarity of those tagged by U3, namely R5, with those already recommended by the system, namely R3. This computing is based on matrix RR = RT x TR:

$sim(R_3,R_5) = cos(RR_3,RR_5) = \frac{(1\ 0) \text{ x } (0\ 1)}{\sqrt{1+0} \text{ x } \sqrt{0+1}} = \frac{0}{1} = 0$

Then R5 and R3 are not similar and R5 is weakly recommended to U1 [Beldjoudi et al., 2011b].

In order to make our approach scale to very large databases by avoiding repeated recalculations, we enrich our dataset with facts extracted from similarities that have been already calculated. These facts assert that a resource X is similar to a resource Y. For example, let us suppose that we want to know if resource Rx is relevant for user U. In this case before going to calculate the similarities between this user and the other taggers who employed this resource, we first search for resources similar to resource Rx, by checking if there exists a triple (Rx, IsSimilarTo, Ry) in the database. In this case our system will not recalculate the similarity between user U and the taggers who used this resource, nor recalculate the similarity between these two resources. It will directly propose resource Ry to U with the same recommendation level of Rx.

Let us note that the choice of this kind of facts was based on resources and not on users because we are aware that user profiles can be changed at any time by adding or removing new tags or new resources and therefore we cannot assert that two users will always have the same tastes. On the contrary if a large set of users has already agreed that two resources are similar, this information becomes an assertion even if the profiles of these users can be changed in the future. And so we can assume that two resources are similar if they have already been judged as similar by an important group of users.

In order to make our proposal more understandable, let us consider the following example: Suppose that two resources R1 and R2 are two chapters about web 2.0. At a given time, 5000 users agree that these two resources are similar: they tagged these resources by common tags.

After a period the profiles of these users have changed (resources and/or tags have been added or removed) and some of them changed their interests. This can affect the similarity value between resources R1 and R2 which becomes lower than the similarity threshold. These resources then become dissimilar in the system which is contradictory because these two chapters treat the same subject. In our approach we represent and save such similarities in order to avoid losing them.

4.2.3 Resolving Language Variations in Recommendation

Multilingualism, dialects and spelling variations are the cause of the most annoying effects in recommender systems. The user perceives the negative effect when the system cannot give him the resources related to a specific tag used in his search.

Recommendation of Similar Resources

Let us consider the illustrative example in figure 4.2. When a user searches for all the resources related to the tag "football", the resources tagged with "foot" and "soccer" will not be proposed to him. In order to show the negative effect of this situation on resource recommendation based on association rules, let us consider for example that the association rule sport $\Rightarrow$ football holds (see figure 4.2).

According to the method described in section 4.2.2, the recommendation system would propose only the resources related to tag "football" to users having tag "sport" in their profiles. The resources tagged with "foot" and "soccer" would not be proposed to this user.

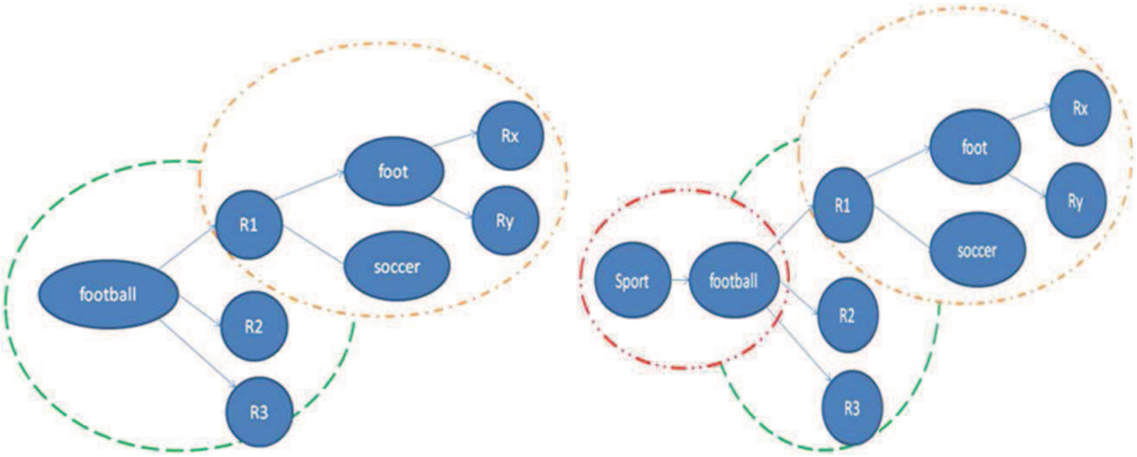


Figure 4.2: An illustrative example with and without an association rule sport ⇒ football

To answer this problem, we introduced the following steps in our process:

-for each user, for each tag found in the consequence of a rule: calculate the similarity between each resource which is tagged by it and is highly recommended and the other resources having another common tag with this recommended resource;

-select the resources which are similar to the first one;

-recommend these resources to the corresponding user with the same level of recommendation.

For instance, in the above example, suppose resource R1 is highly recommended. The process becomes as follows:

-the similarities between R1 and Rx and Ry which are tagged like R1 with "foot" are calculated;

-Rx is selected which is similar to R1;

-Rx is recommended to the user with the same level of recommendation than R1.

Enrichment of the folksonomy

Let us now consider the following situation. Suppose that both the association rule software ⇒ computer and the user profiles in table 4.6 hold.

In this case, according to our approach, resources R3 and R4 will be recommended to user U1, but not resource R6 despite the fact that it seems relevant to U1's preferences. Also resource R7 used by U4 gives the impression that it is significant and adequate to enrich U1's profile even if it is not a language variation of ‘‘computer’’.

Users	Tags	Resources
U1	Software	R1
U1	Software	R2
U2	Computer	R3
U2	Computer	R4
U2	Software	R2
U2	Programming	R1
U3	Java	R5
U3	computer-science	R6
U4	Informatics	R7

Table 4.6: Example of a folksonomy

To answer this problem we enrich the folksonomy by applying the following rule:

If (RX,IsSimilarTo,RY) ∧ (T1,IsSimilarTo,T2) ∧ (RX,TaggedWith,T1) ∧ (RY,TaggedWith,T2) Then (RX,CanBeTaggedWith,T2) ∧ (RY,CanBeTaggedWith,T1)

Let us consider the facts extracted from the above example of a folksonomy:

(R1, TaggedWith, Software) ∧ (R2, TaggedWith, Software) ∧ (R3, TaggedWith, Computer)

(R4, TaggedWith, Computer) ∧ (R2, TaggedWith, Software) ∧ (R1, TaggedWith, Programming)

(R7, TaggedWith, Java) ∧ (R6, TaggedWith, Computer-science) ∧ (R7, TaggedWith, Informatics) and the following association rule: software ⇒ computer.

Let us now suppose that the two facts (R3, IsSimilarTo, R6) and (R3, IsSimilarTo, R7) have been extracted from a previous calculations.

The above rule enables to infer and add the following two facts in the folksonomy:

(R6, CanBeTaggedWith, Computer) and (R7, CanBeTaggedWith, Computer).

Our recommender system will then recommend R6 and R7 to user U1 because it detects that these two resources are relevant to enrich U1's profile.

Thus, we have exploited the strength of social aspect in folksonomies to let each member in the community benefit from the resources tagged by his other neighbors in the social networks based on resources recommendation. Our objective in the next section is to propose a personalized ranking function to rank the retrieved resources.

4.3 Personalized Ranking Function in Collaborative Environment

In folksonomies, a relatively large number of resources can match users' queries. Therefore, ranking these web resources is a key problem, since a user cannot browse all these resources. Ranking in information retrieval consists of the definition of a ranking function that allows quantifying the similarities among documents and queries. A ranking function should incorporate many features to be effective, e.g. features of the document, the query, the overall document collection, the user, etc [Bouadjenek et al., 2013].

In this section, a personalized social ranking function is proposed to provide personalized retrieved resources.

Within the context of folksonomies, we can formalize the ranking problem as follows:

Let's consider a folksonomy F (U, T, and R) from which a user *u*∈ U receives recommended resources or submits a query *t* to a search engine. We would like to rank the set of resources that match *t*, such that relevant resources for *u* are highlighted and pushed to the top for maximizing his satisfaction and personalizing the search results.

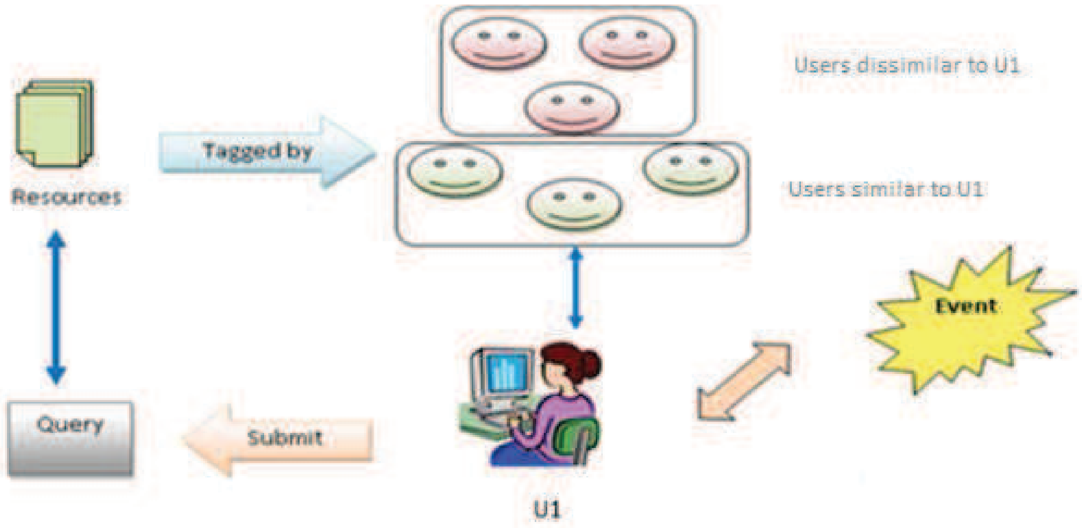


Figure 4.3: Overall of process in collaborative environment

Folksonomies allow users to distribute and receive significant resources about real-world events. This content may appear in various forms, including status updates, photos, and videos, that can be created or posted before, during, and after an event. Furthermore, for known and planned events, structured information (e.g., title, time, location) might be available through event-aggregation social media sites (e.g., Facebook Events, Meetup, EventBrite). Such prior knowledge, however, is not available for unknown or spontaneous events (e.g., natural disasters). By automatically identifying the social media content related to either known or unknown events.

The objective of this approach is improving resources retrieval by exploiting the dependence between event detection and the high number of similar queries transmitted in the same period.

It is clear that with the presence of a particular event, the number of similar queries in search engines increases considerably.

The presence of an event e involves increasing the number of searches performed on this event. Real time event (RTE) is a formula that can detect the presence of an event related to a given query since the increase of similar queries in a particular period [Boughareb and Farah, 2012].

The RTE score is defined by the following equation:

$$RTE = \sqrt[3]{\frac{A+B}{B}} - 1 \qquad (4.2)$$

Where A is the number of related queries in n time units, B is the number of queries in m units during earlier times, such as n≥ 2 and m=n/2.

For example, taking a week as a time unit, if n = 2, the RTE score will determine the increase of the number of queries submitted about a given topic in two weeks.

To achieve a good ranking, we must sort the retrieved resources according to some criterion so that the most relevant results appear early in the retrieved list displayed to the user. Two cases can be observed:

1) There is not an event detected during the recommendation phase (i.e. the RTE score is lower than a defined threshold *z*): in this case the retrieved resources will be ranked only according to social similarities values.

2) Otherwise, when the value of RTE is greater or equal than *z*, the proposed ranking function incorporated two features to be effective: the social aspect and the number of visit of each resource during the same time period of event detection.

Thus the proposed function $Rank\ (r, t, u)$ is computed by merging the average similarities values $\frac{\sum_{i=k}^{j} sim(u1,ui)}{j}$ and $Nbr\_visit\ (r)$. This merge is computed as follows:

$$Rank\ (r, t, u) = \alpha \times \frac{\sum_{i=k}^{j} sim(u1,ui)}{j} + (1 - \alpha) \times Nbr\_visit\ (r) \qquad (4.3)$$

Where, the parameter α denotes the weight that satisfies 0 ≤ α ≤ 1 and represents the importance one wants to give to the two types of features, i.e. social similarities or most popular resource in n time unites.

In fact, depending on the context, one may want to give a higher importance to users' similarities. Another user may want to give more importance to special events that can be occurring, and so prefers the most popular resources.

Note that, the first side of the formula (4.3) represents the average value of similarities between u and *(j)* other users who tagged a given resource (*r*) with the tag (*t*). It is clear that if two users are dissimilar, then similarity between resources replace that between those two users.

The introduction of last feature in the ranking function is crucial, because the presence of an event can have a real influence on the popularity of resources and thus influenced the meaning of tags in the query even if those later are ambiguous.

In order to estimate the utility of taking into account event detection in the ranking function, let us consider the following example where we have four users' profiles described with their tag list as follow:

U1 (TP, Java, Apple); U2 (Eclipse 3.4, TP, Java); U3 (Apple, Java, Programming, Eclipse) and U4 (Eclipse 3.5, Java, TP)

If the system recommend to user U1 resources related to the tag 'Eclipse', all the resources tagged with the tags (Eclipse 3.4, Eclipse, Eclipse 3.5) will be proposed to our user. But the question in which order those resources should be retrieved?

We distinguish between two different cases:

Case 1: if there is not an event presented during this period, our approach was focused only on social similarities (by taking the value of α equal to 1) in order to rank the retrieved resources.

Case 2: suppose that we have the following event: there is a new version of Eclipse in the net. Thus, if U1 wants to obtain resources about the tag 'Eclipse', it appears relevant to him propose firstly the resources tagged with *eclipse 3.5* before those tagged with *eclipse 3.4* and even with *eclipse*.

We conclude that event detection can play a crucial role in ranking retrieved resources within collaborative environments.

4.4 Conclusion

In this chapter we proposed a contribution about resources recommendation and ranking within folksonomies. Firstly; we have exploited the strength of social aspect in folksonomies to let each member in the community benefit from the resources tagged by his other neighbors in the social networks based on resources recommendation. We have seen that it is very important to analyze the users profile in order to realize a dynamic recommendation can be adapted to each modification in the users' interests. Starting from this point, we have found that it is very significant to overcome the semantics problems within folksonomies during our recommendation. The followed method is based on the similarities between users in some cases and between resources in the other cases. We exploited association rules extracted from the social relations in a folksonomy to recommend resources tagged with terms occurring in these rules in the social network. Our objective is to create a consensus among users of a same network in order to teach them how they can organize their web resources in a correct and optimal manner.

In another side, we proposed to take profit of social interactions between users where the objective was the combination between social similarities and event detection to rank retrieved resources. In the following chapter we will test our approach over different datasets to measure its effectiveness.

Chapter 5:
Evaluations and Experimental Results

Chapter 5:

Evaluations and Experimental Results

We presented in the previous chapter approaches that aim to leverage the social dimension of the web for improving the IR process. The current chapter describes details of implementing the proposed approaches in order to evaluate and demonstrate their efficiency. Also, we implemented a real world application for diabetes disease, which put into practice these approaches. This chapter is organized as follows: Section 5.1 describes experiments of resources recommendation over two baseline datasets. In Section 5.2 we will detail the evaluation of resources recommendation over a real world application for diabetes disease and in section 5.3 evaluations of searching and ranking model in collaborative E-learning will be presented. Conclusions are described in Section 5.4.

5.1 Experiments of Resources Recommendation over Baseline Datasets

In order to evaluate the performance of our recommender system, we have conducted an experiment over two baseline datasets in folksonomies field: del.icio.us and Flickr. Both experiments are described and the results are analyzed and discussed.

5.1.1 Del.icio.us database

To validate our approach, we have conducted a first experiment with del.icio.us data. Our test base comprises 58588 tag assignments involving 12780 users, 30500 tags some of which are ambiguous and have many spelling variations, 14390 resources each having possibly several tags and several users. Our system has extracted a set of 946 association rules from the analysis of the dataset with a support equal to 0.5 and a confidence equal to 0.6. We have for example the rule computer ⇒ programming: 60% of the users using the tag "computer" also use the tag "programming".

To demonstrate the validity of our approach, we have distinguished two classes of users: the first one contains the users who employed ambiguous tags and the other one those who did not. The ambiguity of tags has been subjectively decided based on the use context of the tags and their definition in external sources like WordNet. For example the tag "apple" has been used to annotate both the resource *www.nutrition-and-you.com/apple-fruit.html* which is relative to fruits, and other resources like *www.apple.com* that is relative to computers. So we can conclude that the tag "apple" has several meanings, i.e. refer to several concepts and thus is an ambiguous term. On the other hand the users who used the tags "computer", "java" and "programming" are annotating similar web resources, and so we can conclude that these tags are not ambiguous.

5.1.2 Flickr Database

We have conducted a second experiment with the Flickr database. Our test base comprises 37967 tag assignments involving 11567 users, 26876 tags some of which are ambiguous or have spelling variations and 9321 resources: here again each resource possibly has several tags and several users. In this second experience we have also distinguished two classes of users: those who employed ambiguous tags and those who did not. Our system has extracted a set of 476 association rules from the analysis of the dataset with a support equal to 0.5 and a confidence equal to 0.6.

5.1.3 Experimental Methodology

Normally, in order to evaluate the quality of a recommender system, we must demonstrate that the recommended resources are really being accepted and added by the users. Because the knowledge of this information requires asking the users of the selected databases if they appreciated the proposed set of resources, which is impossible in our case because we don't know this community, we have randomly removed some resources from the profile of each user, and we applied our approach on the remainder dataset in order to show if it can recommend the removed resources to their corresponding users or not. If it is the case, so we can conclude that our approach enables to extract the user preferences.

In order to test the performance of our approach we have proposed to follow the following steps:

a) Evaluating the capacity to overcome the ambiguity problem

To achieve this goal, we started by selecting a set of tags containing ambiguous tags; this set consisted of 1154 tags from the del.icio.us database and 563 tags from the Flickr dataset.

Then we have randomly removed sets of resources tagged by these ambiguous tags. Let us note that all the removed resources were randomly selected in order to preserve the justice and the integrity of our evaluation. We repeated this process five times for each tag in order to make a cross-validation. In other words for each tag we have divided its corresponding set of resources randomly to five parts and then selected one part to be removed in each evaluation in order to use it as a test set. This process was repeated five times and in each time we have selected a different test set from the divided parts.

- Experimental Results: In order to evaluate the quality of our recommender system, we have used the following three metrics: recall, precision and F1 metric that is a combination of recall and precision.

Based on our test datasets, we extracted 107 association rules from the die.icio.us dataset and 98 one from the Flickr dataset, with a support equal to 0.5 and a confidence equal to 0.6. Afterwards we calculated the three metrics for each participant in our test. Table 5.1 presents the average values of the metrics.

	Precision	recall	F1
Del.icio.us dataset	77%	83%	80%
Flickr dataset	84%	90%	87%

Table 5.1: Average precision, recall and F1 of the recommendations

These results showed that, by applying the extracted association rules, the resources associated to non ambiguous tags are highly recommended. It has also showed that, in the case of rules involving ambiguous tags, our system recommends to the user the resources which are close to his interests with a high level of recommendation and, on the contrary, those which are far from his interests with a low level of recommendation.

b) Evaluating the capacity to overcome the spelling variations problem

To achieve this second goal, we started by selecting a set of tags containing terms with many spelling variations; this set consisted of 2417 tags from the del.icio.us database and 1186 tags from Flickr dataset. Then we have randomly removed resources tagged with these tags in order to test weather our system recommends them to their right users. We repeated this process five times in order to make a cross-validation.

- Experimental Results: Based on our test datasets, we have extracted 127 and 101 association rules respectively from the del.icio.us and Flickr database, this with a support equal to 0.5 and a confidence equal to 0.6. Afterwards we have calculated the three above metrics for each user. Table 5.2 presents the average values of the three metrics.

	Precision	Recall	F1
Del.icio.us dataset	69%	80%	75%
Flickr dataset	66%	77%	72%

Table 5.2: Average precision, recall and F1 of the recommendations

5.1.4 Discussion

From the analysis of the above results we can conclude that, in all scenarios, precision, recall and F1 of our approach are very promising both in del.icio.us and Flickr datasets. These results indicate that the use of association rules and social similarities performed by our approach are really able to take into account users profiles when recommending resources.

Not surprisingly, our experiment has showed that the resources associated to no ambiguous tags are highly recommended. It has also showed that, in the case of ambiguous tags, our system proposes to the user the resources which are close to his interests with a high level of recommendation and, on the contrary, those which are far from his interests with a low level of recommendation. It has also showed that when a user wants to obtain relevant resources concerning a specific tag, the majority of pertinent resources related to the tags which are spelling variations of the entered one are given to this user.

To sum up, the consensus among users who have similar interests for using the same tags or the same resources plays an important role in the elimination of the ambiguity problem. Also increasing the weights of these tags or these resources makes the semantics emerge even when there are tags that can have several meanings. The results presented in the above tables (table 5.1 and 5.2) show a rate of precision and recall very optimistic in the data set tested in this experience. Indeed these results show that our approach succeeds in distinguishing between ambiguous tags and taking into account spelling variations during the resources recommendation. An analysis of our approver's correctness will be presented in the next subsection.

5.1.5 Analyze of the Approach Accuracy

In order to analyze the accuracy of our approach, we compared our results against the null hypothesis where every resource tagged with an ambiguous tag is returned. We consider a naive folksonomy without any method to overcome the semantics problems between tags. The average rates of precision, recall, and metric F1 obtained are presented in table 5.3.

-Tags Ambiguity: When omitting the steps proposed in our approach, the rates of precision become very low, which confirms that the folksonomy suffers from the precision of results and so the ambiguity problem in the step of resources retrieval, and no respect of users' preferences in the resources recommendation process. Also the metric F1 rate decreases according to the diminution of precision. On the contrary the rates of recall are very high (100%), this can be explained by the ability of our system to retrieve and so recommend all the existing resources by a simple selection query.

-Spelling Variations: When omitting the steps proposed in our approach, the rates of precision becomes very low, which confirms that the folksonomy suffers from the precision of results in the information retrieval and so in resources recommendation. The rates of recall are also much lower than with our approach. This can be explained by the inability of the system to retrieve all the relevant resources tagged with tags related to the one found in the rule consequence. The rate of the metric F1 also decreases.

Problem	Database	Precision	Recall	F1
Tags ambiguity	Del.icio.us	12%	100%	21%
Spelling variations		44%	10%	16%
Tags ambiguity	Flickr	33%	100%	50%
Spelling variations		25%	20%	22%

Table 5.3: The average values of the three metrics concerning the problem of tags' ambiguity and spelling variations without following our proposed approach

To conclude, the values of precision and recall achieved with our approach are very promising. Especially when we consider the F1 metric, we can observe that our approach achieves the best values. This implies that it is the most adequate when the user wants to obtain a trade-off between precision and recall. The use of association rules and social similarities really enable to satisfy the user's need when recommending him a set of resources. Tables 5.4 and 5.5 present the deviation value of precision, recall and the F1 metric in both del.icio.us and Flickr datasets for tags ambiguity and spelling variations problems respectively.

	Precision	Recall	F1
Del.icio.us dataset	5%	6%	5%
Flickr dataset	7%	5%	6%

Table 5.4: The standard deviation value of the three metrics concerning tags ambiguity problem

	Precision	Recall	F1
Del.icio.us dataset	8%	5%	4%
Flickr dataset	9%	4%	5%

Table 5.5: The standard deviation value of the three metrics concerning spelling variations problem

In both cases, these values are very small which indicates that the value of these measures for each user tend to be very close to the average. Since the averages (presented in tables 5.1 and 5.2) are very promising for the community in general, the small values of standard deviations indicate that the metrics are also promising for each user individually.

5.1.6 Choice of the Optimal Value for Support and Confidence

The aim of association rules mining is to find all the rules that satisfy certain minimum support and confidence restrictions. The more we augment the support value, the more the extracted rules are evident, and thus, the less they are helpful for the user. As a result, it is necessary to put the support value low enough in order to extract important information. Unfortunately, when support threshold is very low, the volume of rules becomes very large, making it difficult to analyze the obtained rules.

The confidence is only an estimate of the rules' accuracy in the future. It represents the confidence that we want in the rules.

A certain amount of expertise is needed in order to find the relevant support and confidence settings, to obtain the best rules that impact on the rate of F1 measure.

To find optimal values of minimum support and minimum confidence, two experiments are done. In the first experiment, we search the optimal value of minimum support using the two datasets del.icio.us and Flickr. We choose different value of minimum support ranging from 0.1 to 1 to select the value for which our approach has the best performance. Figure 5.1 shows the F1 metric evolution based on the selected minimum support using the two experimental datasets.

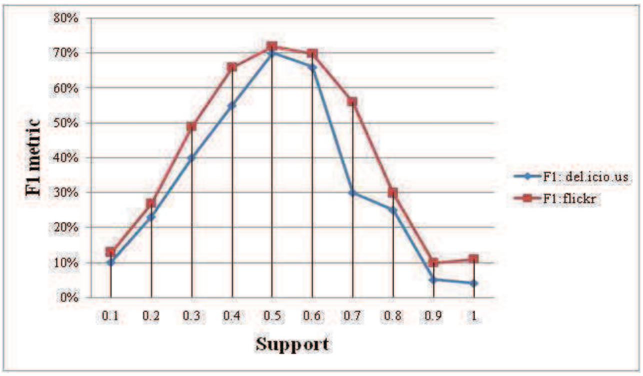


Figure 5.1: Optimal value of minimum support

As can be seen in this figure, the most suitable value of minimum support that produces the highest value of F1 metric is 0.5.

The second experiment concerns the search of the optimal value of minimum confidence using also the two experimental dataset del.icio.us and Flickr where minimum support = 0.5. In this experiment, different values of minimum confidence are used ranging from 0.1 to 1. Figure 5.2 shows the value of F1 metric evolution based on the selected value of minimum confidence.

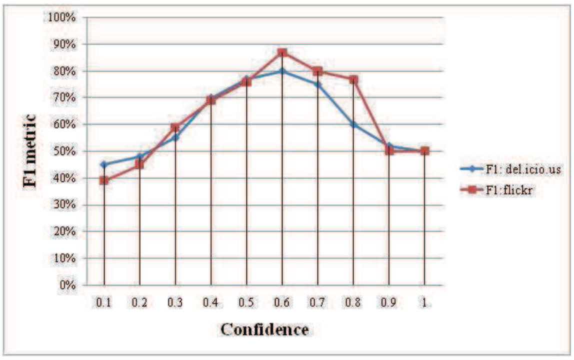


Figure 5.2: Optimal value of minimum confidence

From this figure, the optimal value of minimum confidence that provides the best performance is 0.6.

In the resulting experiments, the relevant support and confidence settings are 0.5 and 0.6 respectively.

5.1.7 Similarity Threshold

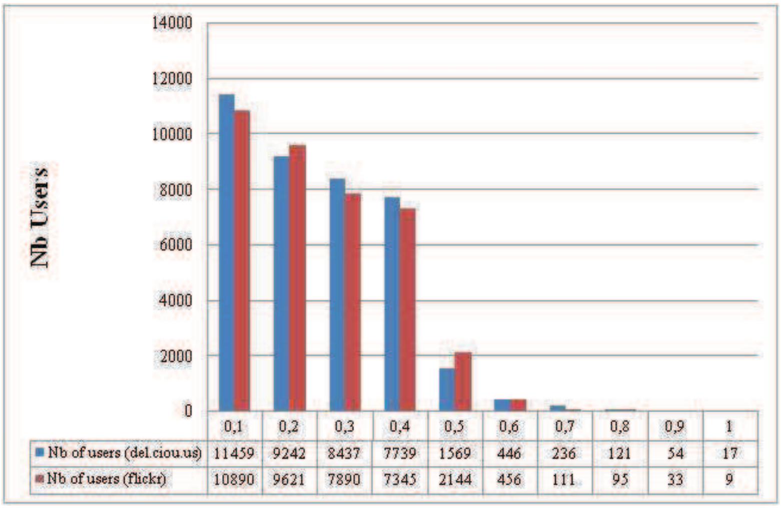
The distribution of tags over resources and users in folksonomies follows a power law [De Meo et al., 2010, Zanardi and Corpa, 2011]: most resources are tagged by only a few numbers of users, and many tags are only used by a few numbers of users. This intensely impacts on the similarity degree between two users. Figure 5.3 shows that almost all pairs of users examined in the experimental datasets (del.icio.us and Flickr) showed a very low similarity degree.

In order to choose the relevant threshold value of similarity among users and among resources, we have selected many thresholds distributed in the interval [0, 1]. In our experiment we remarked that:

-when we choose law values of similarity threshold, our approach generates many incorrect similarity relationships among users and among resources.

-on the other hand, when we choose high values, our approach cannot detect some similarity relationship either among users or among resources.

-Intermediate values let our approach detect many correct similarity relationships and to remove most of the incorrect ones.



	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
■ Nb of users (del.ciou.us)	11459	9242	8437	7739	1569	446	236	121	54	17
■ Nb of users (flickr)	10890	9621	7890	7345	2144	456	111	95	33	9

Figure 5.3: Distribution of users' similarity with cosine formula over many thresholds distributed in the interval [0, 1]

In the literature, we find that most similarity measures are based on set intersection, union and cardinality. These similarity measures range between a minimum and a maximum value. Generally these two values are 0 and 1, i.e. the similarity between two objects X and Y is limited as follows:

$$0 \leq \text{sim}(X,Y) \leq 1 \qquad (5.1)$$

To determine whether two objects are similar or not, we must compare their similarity with a defined threshold. The problem of finding the relevant threshold setting is generally resolved empirically.

We propose a new formula that limits the choice of similarity threshold S during the calculation of similarity within folksonomies.

The idea is to calculate the ratio between the number of common tags between two users and the number of tags used by the user who has the richest profile.

$$\frac{min|Ux \cap Uy|}{max|Uz|} \leq S \leq \frac{max|Ux \cap Uy|}{max|Uz|} \qquad (5.2)$$

Based on the matrix UU:

$\frac{min|Ux \cap Uy|}{max|Uz|}$: is the minimum value found in the matrix without including diagonal.

$\frac{max|Ux \cap Uy|}{max|Uz|}$: is the maximum value found in the matrix without including diagonal.

$max|Uz|$: is the maximum value found in diagonal.

Let us consider matrix UU=UT*TU. It is characterized by the following properties:

-it is a symmetric matrix.

-each cellule in the diagonal represents the number of tags used by user Uz, which gives us |Uz|.

-the values of other cells outside the diagonal represent the number of common tags between two users Ux and Uy (i.e. |Ux ∩ Uy|).

	U1	U2	U3	U4	U5
U1	278	146	0	132	0
U2	146	246	100	0	0
U3	0	100	144	0	44
U4	132	0	0	132	0
U5	0	0	44	0	44

Table 5.6: Example of a matrix UU

$$\frac{min|Ux \cap Uy|}{max|Uz|} = \frac{0}{278} = 0$$

$$\frac{min|Ux \cap Uy|}{max|Uz|} = \frac{132}{278} = 0.47$$

So the similarity threshold should not overstep 0.47 in this folksonomy.

We have empirically determined that the best tradeoff was obtained when the threshold value of similarity among resources is equal to 0.45 and that of similarity among users is equal to 0.5.

5.1.8 The impact of α value in the ranking function

In this experiment, we detected the presence of 43 events by using the formula (4.2). To evaluate the impact of these events on ranking the retrieved resources, we proceeded as follow:

In the same time periods where we have detected the presence of an event, two experimental scenarios related to choosing the value of parameter α were proposed:

a) The α value is between 0 and 0.4 ($\mathbf{0 \leq \alpha \leq 0.4}$)

In this case, we remarked that when the value of α is between 0 and 0.4 the ranking function is focused much more on the second feature of the formula (4.3), which is the popularity of resources when event detection.

Experiments demonstrated that in the case of ambiguous tag or tag with spelling variations, when two different events are detected, precision value is bended down because we have neglected the social similarities between users that can personalized the retrieved results. Thus focusing only on event detection can decrease the pertinence of ranking function.

b) The α value is between 0.5 and 1 ($0.5 \leq \alpha < 1$)

It is clear that in this second case the ranking function focused on the social aspect more than the new event influence on resource popularity.

This can degrade the result precision because even with the fact that users are similar, there are some resources became not relevant to an active user because they are very former and thus their popularity decreased comparing with other new resources that are presented recently with the new event detection.

Figure 5.4 demonstrates that the most suitable value of α that produces the highest value of the average of F1 score and the mean reciprocal rank is 0.6. This implies that it is the most adequate when the user wants to obtain a trade-off between social similarities and new events detection.

Note that mean reciprocal rank is a statistic measure for evaluating any process that produces a list of possible responses to a sample of queries, ordered by probability of correctness. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries Q:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \qquad (5.3)$$

The reciprocal value of the mean reciprocal rank corresponds to the harmonic mean of the ranks.

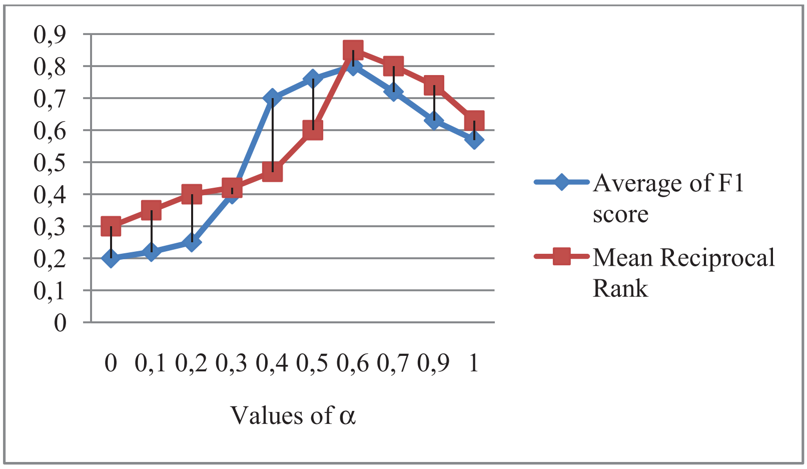


Figure 5.4: Evaluation the impact of α values on average of F1 score and mean reciprocal rank

5.1.9 Scale-up Experiment

As recommender systems are designed to help users navigate in large collections of items, one of our goals is to scale up to real datasets. So, it is important to measure how fast does our approach provides recommendations. In this subsection we discuss the impact of increasing the number of users on the execution time of our approach. In order to demonstrate the scalability of our approach, we measured the execution time required to make relevant recommendations both in del.icio.us and Flickr databases, with a number of users increasing from 1000 to 11500 users.

Figure 5.5 shows that the execution time (in seconds) of our approach linearly increases as the database size increase, meaning that our approach have relatively good scale-up behavior since the increase of the number of users in the database will lead to approximately the linear growth of the processing time, which is desirable in the processing of large databases.

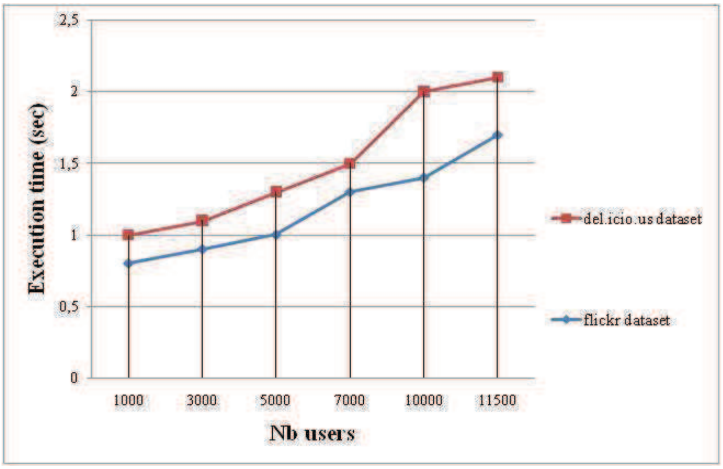


Figure 5.5: Performance of our approach when the database size increases

5.1.10 Comparative Analysis

We propose in this subsection a quantitative comparison between our approach and some approaches for resource recommendation in folksonomies based on the enrichment of the profiles of involved users; in particular, we consider the approaches described in [De Meo et al., 2010], [Huang et al., 2011] and [Zanardi and Corpa, 2011].

These systems show different behaviors; this depends essentially on the different strategies used by them to surmount the folksonomies problems in resources recommendation process. In this context, we are going to analyze each approach from four points (Resolving Tags Ambiguity Problem (RTAP), Resolving Spelling Variations Problem (RSVP), Modeling Users' Preferences (MUP) and Ranking Resources (RR)). In Table 5.7, we report a summarization of three related approaches along with their similarities and differences with ours.

System	RTAP	RSVP	MUP	RR
[De Meo et al., 2010]	No	No	Yes	No
[Huang et al., 2011]	No	No	Yes	No
[Zanardi and Corpa, 2011]	No	No	Yes	Yes
Our approach	Yes	Yes	Yes	Yes

Table 5.7: A comparison between our approach and three related ones

We implemented the three approaches described in chapter 3 and ran both of them and our approach on the del.cio.us dataset described in section 5.1.1. Then we computed the corresponding values of Precision and Recall and F1 metric achieved by each system. At the end of this experiment we averaged the values of the above metrics. In Table 5.8 we reported the obtained results. This table indicates that our approach achieves high values of Precision and Recall and the best value of F1 metric.

Approach	Precision	Recall	F1
[De Meo et al., 2010]	68%	71%	69%
[Huang et al., 2011]	60%	73%	66%
[Zanardi and Corpa, 2011]	72%	60%	65%
Our approach	73%	82%	77%

Table 5.8: Average Precision, Average Recall and Average F1 achieved by our approach and three related ones

In another side, the semantics problems solved in our approach are discussed by previous methods, especially via employing ontologies. In this subsection we will see some comparisons with these methods in order to demonstrate our approach capacity to surmount tags ambiguity and spelling variations when users submit a simple query and not only the final recommendation. In order to make a quantitative comparison between our approach and some approaches aimed to bring together folksonomies and ontologies to overcome the lack of semantics between tags; we considered the approaches described in [Limpens et al., 2010] and [Pan et al., 2010]. In this experiment, we will use del.icio.us dataset described in section 5.1.1.

We implemented the two approaches [Limpens et al., 2010] and [Pan et al., 2010] described in chapter 3 and ran both of them and our approach on the dataset described in section 5.1.1. Next, we have performed some queries for retrieving a set of resources related to a specific tag. This one can be ambiguous and/or have several spelling variations. For each submitted query we computed the corresponding values of Precision, Recall and F1 measure achieved by each system. At the end of this experiment we averaged the values of these metrics across all submitted queries. In Table 5.9 we report the obtained results which indicate that our approach achieves high values of Precision and Recall, and also the best value of F1 metric.

Approach	Precision	Recall	F1
[Pan et al., 2010]	75%	67%	71%
[Limpens et al., 2010]	70%	69%	70%
Our approach	90%	82%	85%

Table 5.9: Average Precision, Average Recall and Average F1 achieved by our approach and two related ones

From the analysis of this table it is possible to observe that the three systems show different behaviors; this depends essentially on the different strategies used by them to surmount the folksonomies problems. In this context, we are going to analyze each approach concerning three points: resolving tags ambiguity problem, folksonomy enrichment and resolving spelling variations problem.

Starting with the approach presented in [Pan et al., 2010] that addressed the problem of tag ambiguity by expanding folksonomy search with ontologies. The author proposed to expand folksonomies in order to avoid bothering users with the rigidity of ontologies. During a keyword-based search of resources, the set of ambiguous used terms is concatenated with other tags so as to increase the precision of the search results. This contribution addresses tags ambiguity problem, however neither the folksonomy enrichment nor the spelling variation problem are tackled. The limits of this approach are listed in the following points: the results of users' queries are not adapted to each user profile; the approach didn't tackle the spelling variations problem and there is no folksonomy enrichment in the proposed method.

In another contribution, [Limpens et al., 2010] focused on using ontologies to extract the semantics between tags. Also, the interactions between users and the system are used to validate or invalidate automatic treatments carried out on tags. The authors have proposed methods to build lightweight ontologies which can be used to suggest terms semantically close during a tag-based search of documents. This work tackled three kinds of relations between tags which are: spelling variations, hyponyms (that include narrower or broader tags) and related tags. The problem of tags ambiguity didn't tackled in this approach, therefore the results obtained when a user wants to search resources annotated by an ambiguous tag can't be personalized according the interest of each user.

Concerning the folksonomy enrichment, we find that the approach of [Limpens et al., 2010] tackled this point by (1) enriching tag-based search results with spelling variants and hyponyms, or (2) suggesting related tags to extend the search, or (3) hierarchically organizing tags to guide novice users in a given domain more efficiently than with at list of tags or occurrence-based tag clouds. The problem of Spelling Variation was tackled in this work, where String-based similarity metrics are applied to tag labels to find spelling variants of tags. The limits of this approach are listed in the following points: the expertise of users that was introduced is characterized by the complexity of its exploitation; the queries results are not adapted to each user profile, also this approach didn't tackle tags ambiguity problem.

In this context, we are going to analyze each approach from the following points (Resolving Tags Ambiguity Problem (RTAP), Resolving Spelling Variations Problem (RSVP), Supporting Related Tags (SRT), Supporting Hyponyms Tags, Supporting Folksonomy Enrichment (SFE), Modeling Users' Preferences (MUP) and Ranking Resources (RR)). In table 5.10 we report a summarization of these two related approaches along with their similarities and differences with ours.

Approach	RTAP	RSVP	SRT	SHT	SFE	MUP	RR
[Pan et al., 2010]	Yes	No	No	No	No	No	No
[Limpens et al., 2010]	No	Yes	Yes	Yes	Yes	No	No
Our approach	Yes	Yes	Yes	No	Yes	Yes	Yes

Table 5.10: A comparison between our approach and two related ones

5.1.11 General Discussion

The results of our experiments are very optimistic and so we can say that the force of community effect in folksonomies applied with association rules have showed its efficiency in the enrichment of users' profiles. At the same time in which our approach contributes to increase the weights associated to the relevant resources, it reduces also tag ambiguity and spelling variations problems. The extraction of association rules is based on tags rather than on resources because we believe that tag popularity in folksonomies is greater than resource popularity and the meaning of tags in these systems is more significant than it of resources. The results presented in the above sections showed rates of precision and recall very optimistic. We must note also that the methodology proposed to treat tags ambiguity and spelling variations problems can be applied during a simple research by tag.

In order to get information about users' feedback on the recommender system, we proposed a real world application for diabetes disease.

5.2 Experiments of Resources Recommendation: A Real World Application for Diabetes Disease

Diseases affect millions of people in the world leading to expensive healthy penalties in our life. Recently with the emergence of social networks and their use in different field, we suggest to exploit this technology in clinical practice by proposing a system based on giving doctors relevant medical resources can be annotated by them. In this process, we propose to take profit from the approach proposed in chapter 4 to show through an empirical scenario, conducted on a real world application for diabetes disease, how we can demonstrate the efficiency of resources recommendation in clinical practice.

Medicine is the science or practice of the diagnosis, treatment, and prevention of disease. It encompasses a variety of health care practices evolved to maintain and restore health by the prevention and treatment of illness.

Medical availability and clinical practice varies across the world due to regional differences in culture and technology. Modern scientific medicine is highly developed in the Western world, while in developing countries such as parts of Africa or Asia, the population may rely more heavily on traditional medicine with limited evidence and efficacy and no required formal training for practitioners. Like most people, healthcare professionals use mainstream social media networks to connect with friends and family. But most of them also join social networks focused exclusively on healthcare. Unfortunately, the current medical social networks still suffer from tag ambiguity and spelling variations.

Clinical Decision Support Systems (CDSS) in health care have a long history where recent analysis show that their number increases. CDSS may be used for different goals such as recommendations, statistical calculations, giving services for reminders or alerts to different user groups, etc [Kaplan, 2001]. Web 2.0 technologies are presented as enablers in health care and clinical practice. In this section, we propose testing the approach presented in chapter 4 within a medical domain. Thus we present a method to analyze doctors' profiles according to their tagging activities in order to predict interesting medical resources and recommend them. The objective is to enrich doctors profiles based on similarities calculation and association rules mining; and by doing so to increase the community effect when suggesting resources to a given doctor.

5.2.1 Resource Recommendation in Medical Social Networks

In modern clinical practice, doctors personally assess patients in order to diagnose, treat, and prevent disease using clinical judgment. The doctor-patient relationship typically begins an interaction with an examination of the patient's medical history and medical record, followed by a medical interview and a physical examination. Basic diagnostic medical devices are typically used. After examination for signs and interviewing for symptoms, the doctor may order medical tests, take a biopsy, or prescribe pharmaceutical drugs or other therapies. Differential diagnosis methods help to rule out conditions based on the information provided. During the encounter, properly informing the patient of all relevant facts is an important part of the relationship and the development of trust. The medical encounter is then documented in the medical record. Follow-ups may be shorter but follow the same general procedure, and specialists follow a similar process. The diagnosis and treatment may take only a few minutes or a few weeks depending upon the complexity of the issue. The components of the medical interview and encounter are:

-Chief complaint: the reason for the current medical visit. These are the symptoms. They are in the patient's own words and are recorded along with the duration of each one.

-History of present illness: the chronological order of events of symptoms and further clarification of each symptom.

-Current activity: occupation, hobbies, what the patient actually does.

-Medications: what drugs the patient takes including prescribed, over-the-counter, and home remedies, as well as alternative and herbal medicines/herbal remedies. Allergies are also recorded.

-Past medical history: concurrent medical problems, past hospitalizations and operations, injuries, past infectious diseases and/or vaccinations, history of known allergies.

-Social history: birthplace, residences, marital history, social and economic status, habits (including diet, medications, tobacco, etc.).

-Family history: listing of diseases in the family that may impact the patient. A family tree is sometimes used.

-Review of systems or systems inquiry: a set of additional questions to ask, which may be missed on the history of present illness: a general enquiry (have you noticed any weight loss, change in sleep quality, fevers, lumps and bumps? etc.), followed by questions on the body's main organ systems (heart, lungs, digestive tract, etc.).

The physical examination is the examination of the patient for medical signs of disease, which are objective and observable, in contrast to symptoms which are volunteered by the patient and not necessarily objectively observable. The healthcare provider uses the senses of sight, hearing, touch, and sometimes smell (e.g., in infection, uremia, diabetic ketoacidosis).

The medical decision-making process involves analysis and synthesis of all the above data to come up with a list of possible diagnoses (the differential diagnoses), along with an idea of what needs to be done to obtain a definitive diagnosis that would explain the patient's problem. On subsequent visits, the process may be repeated in an abbreviated manner to obtain any new history, symptoms, physical findings, and lab or imaging results or specialist consultations.

We intend in this contribution to see the impact of social web applications such as folksonomies in the assistance of physicians during their daily work with the enrichment of their profiles with medical resources seems interesting for them. Our aim is to move towards a general consensus among doctors in medical systems.

The proposed approach represents a social application intended for all doctors with their different level of specialization and expertise. We argue that the automatic sharing of resources strengthens social links among actors, and we exploit this idea to assemble doctors around a single web application to let each one of them benefit from the others' knowledge and expertise.

We begin by giving an overview about the general design of the proposed application:

As presented in figure 5.6, the procedure begins with the arrival of a new patient to a doctor. This latter is embedded in a social network composed of other colleagues with different expertise levels. In this network, the doctors can use a set of tags and medical resources related to their own knowledge and their different studied clinical cases.

Here, when the doctor begins his work, usually he asked the patient some questions about the reason for the current medical visit, the chronological order of events of symptoms, what drugs the patient takes, etc. Then he saves this information in the patient's profile by tagging resources related to his symptoms, historic, etc. This procedure is repeated each time when a patient comes to a consultation.

Like described in chapter 4, Based on this social network and association rules mining, the system will extract a set of rules connecting tags that are frequently appear together. These rules are of the form T1 ⇒ T2 indicate that often when a doctor uses the tag T1, he also uses the tag T2. The usefulness of these rules appears in the following phases:

- Helping doctors to find the appropriate questions in the medical interview.

- Helping doctors to make an appropriate diagnostic.

- Helping doctors to propose a best treatment (details will be cited in the next sections).

Once these rules are extracted, the system will recommend the resources tagged with the tag T2 to doctors who have used only the tag T1.

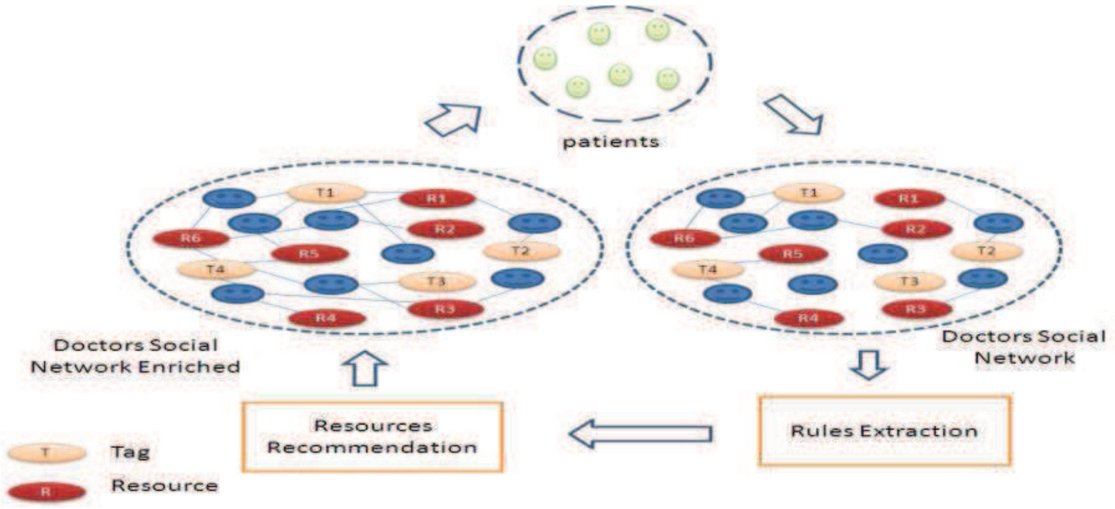


Figure 5.6. Overview of the system design

The aim of this recommendation is to enrich doctors' profiles with relevant medical resources and thus enhancing the social network with other pertinent links between doctors, tags and resources. This enhancement will help doctors to make an appropriate decision to treat their patients' cases. In the next subsection, we propose an experiment describes a real world application carrying novel ideas about diabetes disease. Results will be analyzed and discussed.

5.2.2 A Real World Application for Diabetes Disease

Diabetes affects millions of people in the world leading to substantial negative effects and expensive healthy penalties in our life. Recently with the emergence of social networks in the internet and their use in different field, we propose to use this technology in clinical practice by showing a system based on giving doctors relevant medical resources can be annotated by them. A novel technique is proposed to help doctors discovering the best practices to patient diseases diagnosis and treatments in their daily tasks by analyzing doctors' profiles according to their tagging activity in order to personalize a greatest medical resources recommendation related to the patients' diseases, treatments or clinical cases. We propose to take profit of community effect strength which characterizes social networks with creating and observing emergence of the intelligence captured from socials interactions between doctors in the network by using a powerful method of data mining which is associations rules. We show through an empirical scenario how we can evaluate and demonstrate the efficiency of the medical resource recommender system in clinical decision.

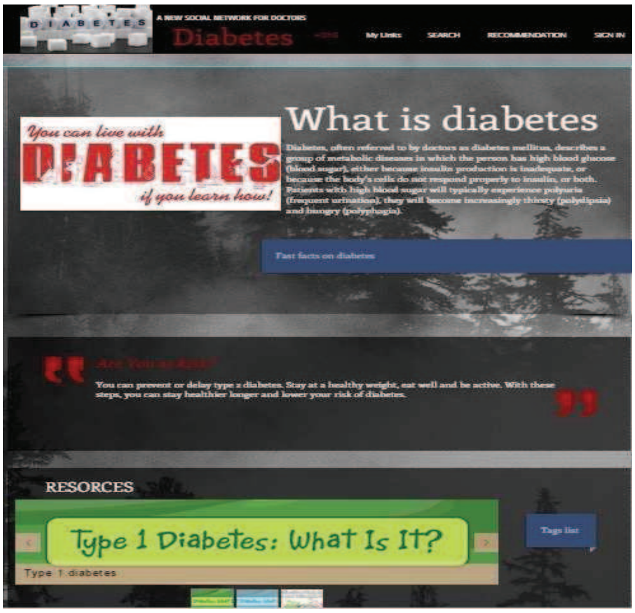
This choice is motivated by the necessity to avoid the problem of knowledge acquisition that gene developers of expert systems since it is relevant to use online knowledge and online community as a potential source of knowledge and moreover enhance traditional explanation and justification with hyperlinks to other relevant web resources.

Because the calculation of the three metrics (Precision, Recall, and the metric F1) requires the knowledge of all relevant resources for each user in order to compare the results provided by our recommender system and those which are preferred by each user, we have built a real database by inviting a set of users to participate in our experiment. Figures 5.7 and 5.8 show two pages in our application.

We have chosen the diabetes disease as subject of this application, this latter is a group of metabolic diseases in which a person has high blood sugar.

Diabetes is no stranger to the 39.5 million people of Algeria, the largest country on the Mediterranean Sea. There are approximately 3 million people with diabetes in this country, part of a sobering reality in the region. According to the International Diabetes Federation, North Africa and the Middle East have the second highest diabetes prevalence rate (9.3%) after North America. Tackling the diabetes challenge in Algeria is complicated by low rates of diagnosis and treatment-typical for many developing countries, resulting in poor gylcaemic control[1].

[1] http://annualreport.novonordisk.com/ [Retrieved 1 May 2015]



What is diabetes

Diabetes, often referred to by doctors as diabetes mellitus, describes a group of metabolic diseases in which the person has high blood glucose (blood sugar), either because insulin production is inadequate, or because the body's cells do not respond properly to insulin, or both. Patients with high blood sugar will typically experience polyuria (frequent urination), they will become increasingly thirsty (polydipsia) and hungry (polyphagia).

Are You at Risk?

You can prevent or delay type 2 diabetes. Stay at a healthy weight, eat well and be active. With these steps, you can stay healthier longer and lower your risk of diabetes.

Figure 5.7: The home page of our application

5.2.3 Dataset

Because our application is incorporated within a web 2.0 technology which is folksonomies, we must give an overview about its three main elements: Resources, Users and Tags.

-Resources: Firstly we made a prototype of a folksonomy in the form of a website, where we have collected a set of different kinds of resources related to the diabetes disease. This set of resources was varied between a set of web pages containing a simple text, videos, photos, …etc. 143 is the number of the collected resources.

-Participants (Users): We have recruited 65 individuals to participate in our study. All participants are doctors interesting with the diabetes disease. The grade of each doctor is varied between internal, general doctor, resident and specialist. We must note that the users of this system are only doctors, and patients have no involvement in this application except through their therapists. All these members are asked to use our real world application in order to show the impact of social interactions in helping each doctor to benefit from the expertise of the others and so let the system move toward a general consensus of its members.

-Tags: The tagging activity is conducted as follow: We have initially asked the specialist doctors to tag a set of resources found on our website in order to let our system benefit from their expertise, and then the use of the system can be done in parallel either by specialists, residents, general practitioners or interns. The number of collected tags is 183 tags.

5.2.4 Experimental Methodology

We have invited a doctors group specializing in diabetes field for participating in our experience. We have initially asked the specialists' doctors to tag a set of resources found in our website, and after that the use of the system can be done in parallel either by specialists, residents, general practitioners or interns. All this to let the non-specialist physicians benefit from the specialists experience within a web 2.0 application like folksonomies.

The profile of each physician is constructed from the set of tags and resources used by him when he treats his patients.

Now in order to link the usual task of doctors to our application, we proceeded as follows: When the doctor begins his work, usually he asked his patient some questions about his symptoms, if he takes already some medicaments, if he suffers from a parallel disease, etc. Then he saves this information in his patient's profile by tagging resources related to his symptoms, his historic, etc. Next, the doctor will make a diagnosis in order to identify the illness of his patient and then save this information in the system in the form of tags linked to resources related to this disease. After the diagnostic phase, it is now the time of therapy. Of course, the treatment proposed by the doctor will differ according to each patient's case. It is the expertise of physician which will be intervened here for proposing the appropriate treatment according to each case. Also this treatment must be cited in the application in the form of tags related to a set of resources indicating information about the proposed medicaments. And of course, this scenario will be repeated each time when a doctor will do a new consultation with one of his patients.

Always we must insist on one of the most strong points in our application which is the dynamism aspect of the users' profiles. Whereas the profile of each patient can be changed in each new consultation: by removing or adding some symptoms, changing one or more therapies, etc. The same, the physician profiles will also be modified according to the arrival of a new patient. Therefore we can say that our system can react according to these updates by adding or removing new tags and new resources. In the next subsection, we will give an overview about the impact of our application on the professional task of doctors:

a) Helping doctors to find the appropriate questions during their questionnaire

Since the specialists' expertise in choosing the relevant questions posed to a patient will be saved in our system, another doctor can benefit from this experience by providing him information (in the form of recommended resources) about the questions or symptoms that he can ask his patients about them in order to discover the correct diagnosis. For example, if our system discovered that the majority of doctors whom found that their patients suffered from the symptoms X and Y they asked them if they suffer also from the symptom Z, then as result the system will generate an association rule X, Y $\Rightarrow$ Z. With this association rule our application will recommend the resources tagged by the tag Z to the doctors whom detected that their patients are suffer from the symptoms X and Y. All this for helping doctors to gather all the necessary information required to make a right diagnostic.

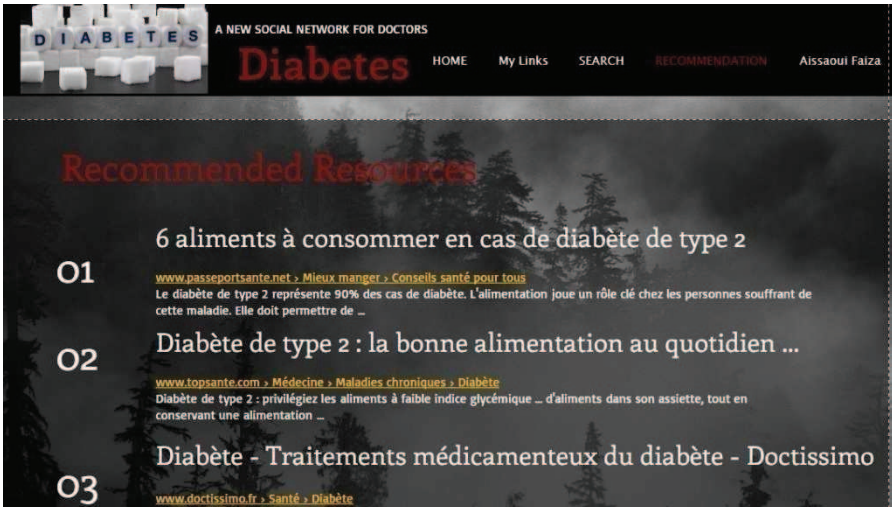
b) Helping doctors to making an appropriate diagnostic

After the questionnaire phase, the doctor arrives at a stage where he must make a correct diagnosis in order to discover the illness of his patient. Similarly, here too our system will greatly help any doctor: an internal, general practitioner, resident and even a specialist for discovering the patient disease focusing on the previous physicians' expertise who have treated similar cases. For example, if the system perceived that the majority of doctors whom detected that their patients suffered from the symptoms X, Y and Z, they diagnosed the disease D, so the system will generate an association rule X, Y, Z $\Rightarrow$ D.

Now, when a new doctor detects that his patient suffers from the symptoms X, Y and Z and he think about the corresponding disease, our system is going to him provide resources related to the illness D for helping this doctor to make an appropriate diagnosis by giving him interesting information about this disease in the form of relevant resources.

c) Helping doctors proposing a best treatment

We are now at the stage where the doctor must propose to his patient the best possible treatment according to his case. In this step each physician must take in the account not only the symptoms from which the patient suffer, but also other considerations such as if the patient take another treatment, if the patient (if she is a women) is pregnant, etc. All this is to avoid proposing a bad treatment for the patients. Here, the strength of our approach will be involved to help the doctors to provide the right treatment for their patients. For example, if the majority of doctors provide the medicine M when they detect the symptoms X, Y and Z, then our system will generate an association rule X, Y, Z $\Rightarrow$ M and therefore it will offer to the physicians who discovered these symptoms, and which are going to seek the appropriate type of treatment, the resources related to the medicine M in order to give them a quick reminder about this remedy and at the same time helping them to propose an appropriate medicament for their patient.



Recommended Resources

O1 6 aliments à consommer en cas de diabète de type 2

www.passeportsante.net › Mieux manger › Conseils santé pour tous

Le diabète de type 2 représente 90% des cas de diabète. L'alimentation joue un rôle clé chez les personnes souffrant de cette maladie. Elle doit permettre de ...

O2 Diabète de type 2 : la bonne alimentation au quotidien ...

www.topsante.com › Médecine › Maladies chroniques › Diabète

Diabète de type 2 : privilégiez les aliments à faible indice glycémique ... d'aliments dans son assiette, tout en conservant une alimentation ...

O3 Diabète - Traitements médicamenteux du diabète - Doctissimo

www.doctissimo.fr › Santé › Diabète

Figure 5.8: Example of resources recommendation for a doctor

5.2.5 Experimental Results

In order to validate our approach efficiency, we propose two experimental scenarios: In the first one we will incorporate the doctors' community in the evaluation process since the calculation of the metrics which will be used in the estimation requires the knowledge of all the relevant resources for each user in order to compare them with the results provided by our recommender system. In the second scenario, we will try to test our approach capacity without involving the doctors in this task. More details will be given in the next subsections.

a) The Fist Scenario:

The aim of this experiment is seeing the impact of association rules in medical field to make a new recommender system.

In order to evaluate the performance of this technique, we choose to calculate the rates of the three metrics Precision, Recall and the metric F1. Based on our test dataset, we have extracted 114 association rules with a support equal to 0.5 and a confidence equal to 0.6. Afterwards we have calculated the above three metrics for each physician. Table 5.11 presents the average values of the metrics we obtained for our 65 doctors.

	Precision	Recall	F1
Average	83%	81%	82%

Table 5.11: The average values of the three metrics following our proposed approach

These are quite encouraging results, showing that our approach of recommendation adapted to doctor profiles is truly able to help doctors when searching for resources.

In order to give an efficient analysis to the obtained results, we have chosen to evaluate our experimental dataset without following the steps and the hypotheses proposed in our approach (presented in chapter 4), and then calculate the rates of precision, recall, and the metric F1. After this evaluation we have obtained the results presented in table 5.12.

	Precision	Recall	F1
Average	17%	100%	29%

Table 5.12: The average values of the three metrics without following our proposed approach

As we see in this table, when we have omitted the steps proposed in our approach, the rates of precision became very low, which confirm that the current folksonomies suffer from the precision of results in the information retrieval because they can't surmount the problem of tags ambiguity, spelling variations and the semantics lack between terms. Also the rate of the metric F1 is decreased according to the diminution of precision. In the contrary the rates of recall show a complete degree (with 100%) which demonstrates the ability of the system to retrieve all the existing resources by a simple select query.

To conclude, the values of precision and recall achieved with our approach are very promising. Especially when we consider the F1 metric, we can observe that our approach achieves the best values. This implies that it is the most adequate when the user wants to obtain a trade-off between precision and recall. The use of association rules and social similarities really enable to satisfy the doctor's need when recommending him a set of resources. Table 5.13 presents the deviation value of precision, recall and the F1 metric in our dataset.

	Precision	Recall	F1
Deviation	6.5%	8%	7%

Table 5.13: The standard deviation value of the three metrics

The presented values are very small which indicates that the value of these measures for each doctor tends to be very close to the average. Since the average is very promising for the community in general, the small values of standard deviation indicate that the metrics are also promising for each user individually.

b) The Second Scenario:

To evaluate the efficiency of our approach without involving doctors in this procedure, we followed the next scenario: we have selected to remove randomly some resources from the profile of each doctor, and then applied our approach on the remainder dataset in order to show if it can recommend the removed resources to their corresponding users. If it is the case, so we can say that our approach can really analyze the doctors' preferences. In order to test the performance of our approach, we propose the following experimental protocols:

- Evaluating the approach capacity to overcome the ambiguity problem in recommendation:

We started by selecting a set of ambiguous tags; this test set consisted of 15 tags. Then we removed random resources tagged with these ambiguous tags in order to see if our approach will be able to overcome the ambiguity problem in its recommendation and recommend the removed resources to their corresponding doctors.

In order to evaluate the efficiency of our recommender system, we used the above metrics: recall, precision and F1 measure. Table 5.14 presents the average values of the metrics:

	Precision	Recall	F1
Average	84%	90%	87%

Table 5.14: The average values of the three metrics concerning the problem of tags Ambiguity

Not surprisingly, our experiment has showed that the resources associated to non ambiguous tags are highly recommended. It has also showed that, in the case of rules involving ambiguous tags, our system recommends to the doctor the resources which are close to his interests with a high level of recommendation and, on the contrary, those which are far from his interests with a low level of recommendation.

- Evaluating the approach capacity to overcome the spelling variations problem in recommendation:

To demonstrate our approach capacity to surmount spelling variations problem, we have also started by selecting a set of tags which have many spelling variations, this set consisted of 35 tags. Then we have removed random resources from these tags in order to judge if our approach will be able to overcome the problem of spelling variations in its recommendation. Table 5.15 presents the average values of the used metrics. These are quite encouraging results, showing that our approach of recommendation adapted to doctor profiles is truly able to help users when searching for medical resources.

	Precision	Recall	F1
Average	69%	80%	75%

Table 5.15: The average values of the three metrics concerning the problem of spelling variations

5.2.6 General Discussion

The obtained results are very promising and so we can say that the force of community effect in folksonomies applied with association rules have showed its efficiency in recommending medical resources to enrich doctors' profiles. At the same time in which our approach contributes to increase the weights associated to the relevant resources, it reduces also tag ambiguity and spelling variations problems. The results presented in the above sections showed rates of precision and recall very optimistic. We must note also that the methodology proposed to treat tags ambiguity and spelling variations problems can be applied during a simple research by tag.

In this section we have exploited the strength of social aspect in doctors' community to let each doctors benefit from the resources tagged by his other neighbors in the social networks based on resources recommendation. We have tested our approach on a real world application for diabetes disease where we have obtained promising results.

To sum up, in the above sections we have demonstrate the efficacy of our approach in recommendation process. Now in order to test its impact within a search procedure, we propose in the following section a motivating example and a scenario to overcome folksonomy problem and obtain relevant ranked resources within a search process without using association rules. Experiments are carried out within E-learning domain.

5.3 Evaluation of Searching and Ranking Model in Collaborative E-learning

Querying a folksonomy is an effective approach that directly retrieves documents from an index of millions of documents in a fraction of a second. Recently, E-learning platforms focused on personalization to achieve pedagogic scenarios otherwise inaccessible to traditional forms of learning. In this section, we propose to evaluate and orient our approach to personalize the resources suggested to each learner when he/she searches relevant resources by using tags.

An experiment over a popular dataset is described to test our approach where results are analyzed and discussed. The dataset exploited in our test is del.icio.us: a web-based social bookmarking tool which allows a user to manage a personal collection of links to web sites and to annotate these links with one or more tags.

In this experiment we were interested with data sample constructed from users who tagged resources about education. Thus our data base comprises 20432 tag assignments involving 7898 users, 15439 tags some of which are ambiguous, 10527 resources each having possibly several tags and several users. Note that, the used dataset include also the date of each tagging operation, this can help us in event detection.

5.3.1 Motivating Example

In social networks, especially in folksonomies, users face some bothered problems like tags ambiguity, synonymy and the lack of semantic links between free used tags. These problems are findable in E-learning folksonomy. For example, in an E-learning application, a learner interested with computer-science can search relevant resources about Sun Company by using the tag (Sun). Unfortunately, in this case the system will him propose not only the resources related to his preferences but also those related to biological element (Sun) since *sun* is an ambiguous tag and such tag is used by learners interested with biology field.

Furthermore, in the case of resources tagged with the same tag, the system cannot access to their content and meaning to differentiate between them, and thus it cannot personalize the retrieved resources according the learner's interests. For instance, two learners in two different departments can study the same course “Algorithmic”, but the course is differed in its content and difficulty from one specialty to another. A learner concerned by computer science is interested with a high level of knowledge about the content of retrieved resources, contrary to a learner in mathematic department. It is clear that the two contents have not the same difficulty, and thus the first learner hasn’t the same feedback about the retrieved resources comparatively to the second one.

In this section, we want to see the impact of searching relevant resources within collaborative E-learning. In figure 5.9, we have four learners and their tags’ list:

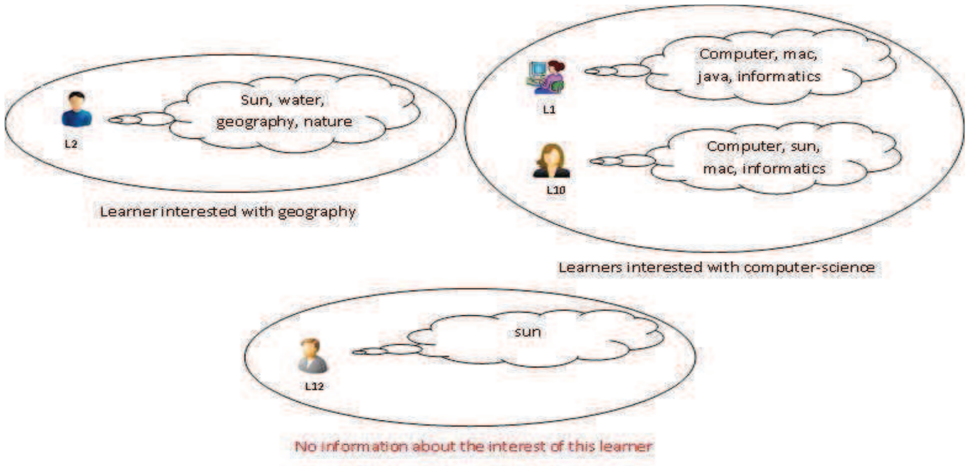

Sun, water,
geography, nature
L2
Learner interested with geography
Computer, mac,
java, informatics
L1
Computer, sun,
mac, informatics
L10
Learners interested with computer-science
sun
L12
No information about the interest of this learner

Figure 5.9: A set of learners with their tags

Let us suppose that learner L_1 wants to retrieve resources related to the tag *'sun'*. In the current folksonomies, the obtained results will contain all the resources tagged with *'Sun'* i.e. those relative to geography and computer-science, even that it is clear to a human reading L_1’s tags, that his preferences are relative to computer and not to geography.

In this section, we demonstrate that by exploiting the proposed approach, if the learner L_1 searches resources tagged with 'Sun'; the system will first propose him the resource corresponding to the tag 'Sun' used by the learner L_{10} with a 'very strong' advice level because the two learners L_1 and L_{10} have similar preferences. On the contrary the resources corresponding to the tag 'Sun' used by the learner L_2 will be given to L_1 with a percentage 'low' of advice because L_1 and L_2 do not share the same interests.

Now in the case of L_{12} for whom not much information about his interests is available, the approach propose to measure the similarity between the resources corresponding to the tag 'Sun' used by L_{12} and the already proposed resources to L_1 with a high percentage (i.e. those of $L_{10)}$. If the resources are similar, the system will propose them to L_1 with a 'very strong' advice level, otherwise with a 'low' advice level.

Let us present the following illustrative example to give more details about the search process.

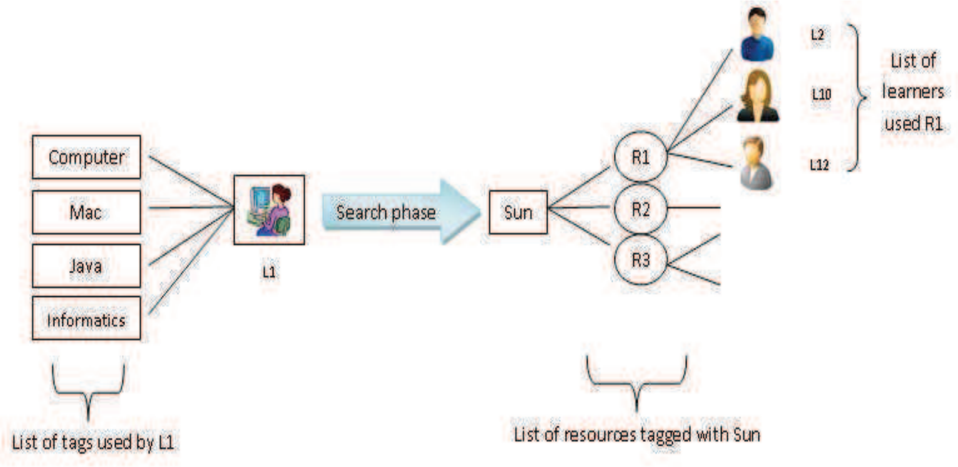


Figure 5.10: An illustrative example

As we see in figure 5.10, the learner L1 want to obtain relevant resources (courses, video, etc.) related to 'sun'. It is clear that before seeing the profile of this learner, we cannot know his preferences because the used tag is ambiguous. For a human reading L_1's tags, we can conclude that L1's preferences are relative to computer and not to geography.

To let the machine understand this, the approach proposes the following steps:

First, construct the matrix RL for the tag 'sun', in which we find the resources tagged with sun and by who.

Second, construct the matrix LT that constitute the profile of each learner in the matrix RL.

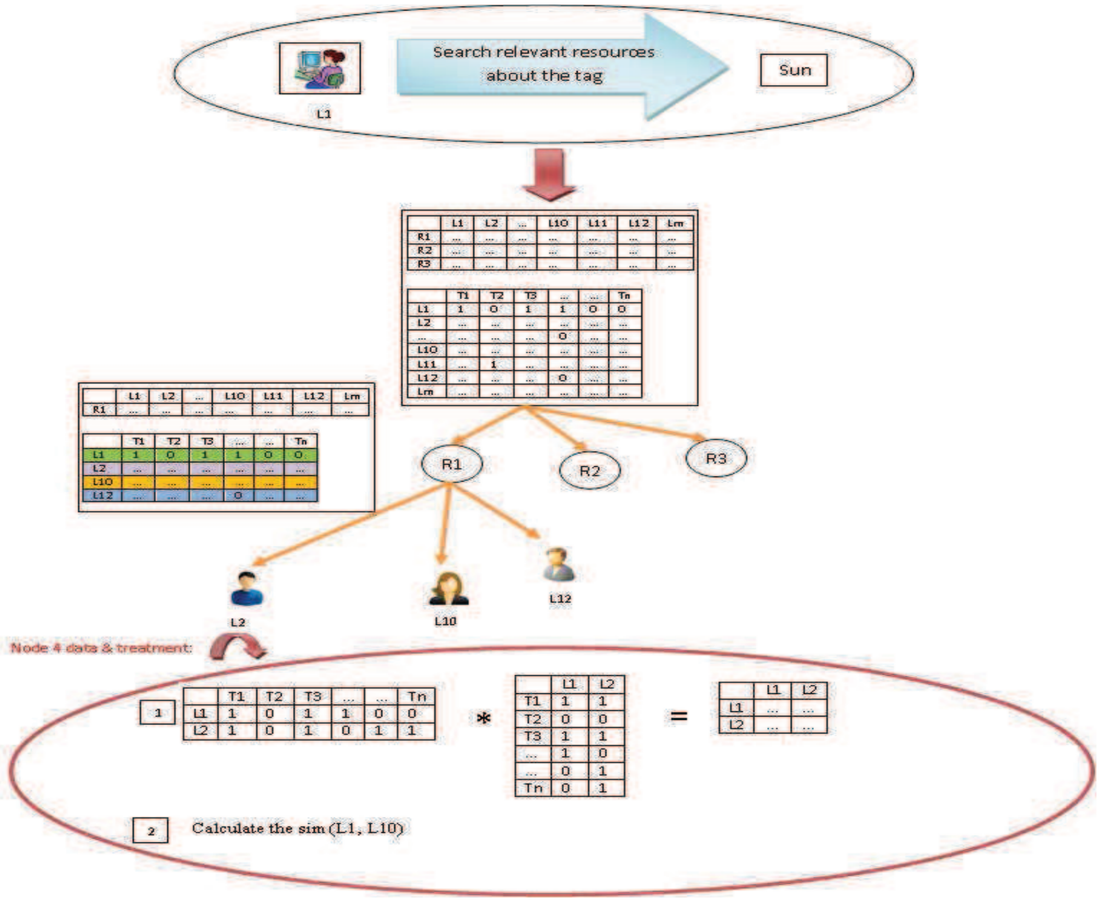
It appears in the matrix RL that the resources tagged with 'sun' are R1, R2 and R3.

Our objective is proposing to L1 (the learner who makes the search) among these resources those close to his preferences. To explain the followed process, we suggest conducted a detailed calculation on R1, and the process will be the same concerning the two other resources R2 and R3.

We interested by the data block extracted from the two matrices RL and LT, especially the row related to R1 in the matrix RL and the rows of LT corresponding to L1 in addition to each user tagged the resource R1 (who are L2, L10 and L12 for example) as is shown in figure 5.11.

To decide if R1 is relevant to L1, we propose to calculate the similarity between L1 and these users. At this stage, we was interested by three sub-blocks because we have 3 users (L2, L10 and L12), in each sub-block we have sub-matrices constructed from two rows of LT one of L1 and the other one from user among those used R1. We need also the transposed matrix (TL) of this later in order to calculate the matrix LL= LT* TL.

Thus, to calculate the similarity between two learners, for example L_1 and L_2, we calculate the cosines of the angle between their associated vectors v_1 and v_2 (designates a series of numbers defined the set of learners' tags).



	L1	L2	...	L10	L11	L12	Lm
R1	...	...	...	...	...	...	...
R2	...	...	...	...	...	...	...
R3	...	...	...	...	...	...	...

	T1	T2	T3	...	...	Tn
L1	1	0	1	1	0	0
L2	...	...	...	...	...	...
...	...	...	...	0	...	...
L10	...	...	...	...	...	...
L11	...	1	...	...	...	...
L12	...	...	...	0	...	...
Lm	...	...	...	...	...	...

	L1	L2	...	L10	L11	L12	Lm
R1	...	...	...	...	...	...	...

	T1	T2	T3	...	...	Tn
L1	1	0	1	1	0	0
L2	...	...	...	...	...	...
L10	...	...	...	...	...	...
L12	...	...	...	0	...	...

1

	T1	T2	T3	...	...	Tn
L1	1	0	1	1	0	0
L2	1	0	1	0	1	1

	L1	L2
T1	1	1
T2	0	0
T3	1	1
...	1	0
...	0	1
Tn	0	1

	L1	L2
L1	...	...
L2	...	...

2 Calculate the sim (L1, L10)

Figure 5.11: The phase of similarities calculation

After similarity calculation between L1 and the users who tagged R1, we calculated the average value of the above similarities to obtain the final value that will be compared with a defined threshold S. The rest of calculations are the same like explained in chapter 4 including the ranking function.

5.3.2 Evaluation Methodology

Making evaluations for personalized search is a challenge since relevance judgments can only be assessed by end-users [Bischoff et al., 2008]. This is difficult to achieve at a large scale. However, different efforts [Bischoff et al., 2008, Krause et al., 2008] state that the tagging behavior of a user of a folksonomy closely reflects his behavior of search on the web. In other words, if a user tags a resource *r* with a tag *t*, he will choose to access the resource *r* if it appears in the result obtained by submitting *t* as query to the search engine.

Thus, we can easily state that any bookmark *(u, t, r)* that represents a user *u* who tagged a resource *r* with tag *t*, can be used as a test query for evaluations.

The main idea of these experiments is based on the following assumption [Bouadjenek et al., 2013]:

For a query $q = \{t\}$ issued by learner (user) *l* with query term *t*, relevant resources are those tagged by *l* with *t*.

Hence, we randomly select 500 pairs *(l, t)*, which are considered to form a personalized query set. For each corresponding pair *(l, t)*, we remove all the bookmarks $(l, t, r) \in F$, $\forall r \in R$ in order to not promote the resource *r* in the results obtained by submitting *t* as a query in our approach. By removing these bookmarks, the results should not be biased in favor of resources that simply are tagged with query terms and making comparisons to the baseline uniformly. Hence, for each pair, the learner *l* sends the query $q = \{t\}$ to the system. Then, we retrieve and rank all the resources that match this query. Finally, according to the previous assumption, we compute the average precision and recall over the 500 queries.

5.3.3 Experimental Results

In order to evaluate the quality of the approach, we used the metrics: recall, precision and F1 metric. The three metrics are calculated for each user, and then the average of each metric is calculated. The results are shown in the table 5.16:

	Precision	Recall	F1
Ambiguous Tags	85%	80%	83%
Not ambiguous Tags	93%	85%	89%

Table 5.16: The average value of the three metrics

From the analysis of the above table we can conclude that, in all scenarios, the Precision, Recall and the metric F1 of our approach are very promising both in the case of ambiguous tag-based queries and also not ambiguous tag-based queries. This result indicates that social similarities performed by our approach are really able to help users when they query a folksonomy.

Not surprisingly, our experiment has showed that the resources associated to no ambiguous tags are highly proposed. It has also showed that, in the case of ambiguous tags, our system proposes to the user the resources which are close to his interests with a high level of suggestion and, on the contrary, those which are far from his interests with a low level of suggestion.

The results presented in the table 5.16 show a rate of precision and recall very optimistic seeing the dataset tested in this experience. Indeed our approach is succeeded in distinguishing between ambiguous tags.

5.3.4 Analyze the Approach Accuracy

In order to analyze the accuracy of our approach, we compared our results against the null hypothesis where every resource tagged with an ambiguous tag is returned. We consider a naive folksonomy without any method to overcome the semantics problems between tags. The average rates of precision, recall, and metric F1 obtained are presented in table 5.17.

Tags Ambiguity: When omitting the steps proposed in our approach, the rates of precision become very low, which confirms that the folksonomy suffers from the precision of results and so the ambiguity problem in the step of resources retrieval, and no respect of users' preferences in the resources retrieval process. Also the metric F1 rate decreases according to the diminution of precision. On the contrary the rates of recall are very high (100%), this can be explained by the ability of the system to retrieve all the existing resources by a simple selection query.

	Precision	Recall	F1
Tags ambiguity	10%	100%	18%

Table 5.17: The average values of the three metrics concerning the problem of tags' ambiguity without following our proposed approach

To conclude, the values of precision and recall achieved with our approach are very promising. Especially when we consider the F1 metric, we can observe that our approach achieves the best values. This implies that it is the most adequate when the user wants to obtain a trade-off between precision and recall. The use of social similarities really enables to satisfy the user's need when retrieving him a set of resources.

Table 5.18 presents the deviation value of precision, recall and the F1 metric in del.icio.us datasets for tags ambiguity problem.

	Precision	Recall	F1
Del.icio.us	6%	8%	7%

Table 5.18: The standard deviation value of the three metrics concerning tags ambiguity problem

These values are very small which indicates that the value of these measures for each user tend to be very close to the average. Since the averages (presented in table 5.16) are very promising for the community in general, the small values of standard deviations indicate that the metrics are also promising for each user individually.

5.3.5 Scalability evaluation

As information retrieval systems are designed to help users navigate in large collections of resources, one of our goals is to scale up to real datasets. So, it is important to measure how fast does our approach provides results.

In this subsection we discuss the impact of increasing the number of learners on the execution time of our approach.

In order to demonstrate the scalability of our approach, we measured the execution time required to make relevant retrieval and ranking in del.icio.us database, with a number of learners increasing from 2000 to 7000.

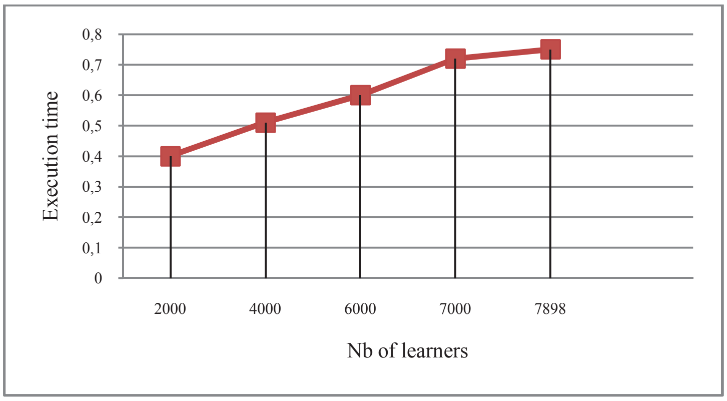


Figure 5.12: Evaluation Performance of our approach when the database size increases

Figure 5.12 shows that the execution time of the approach linearly increases as the database size increase, meaning that our approach have relatively good scale-up behavior since the increase of the number of learners in the database will lead to approximately the linear growth of the processing time, which is desirable in the processing of large databases.

5.3.6 General Discussion

The proposed approach was based on social interactions between learners where the objective was the combination between social similarities and event detection to retrieve and rank relevant resources. The values of precision and recall achieved with our approach are very promising. Especially when we consider the F1 metric, we can observe that our approach achieves the best values. This implies that it is the most adequate when the user wants to obtain a trade-off between precision and recall. Also the deviation values of three metrics are very small which indicates that the value of these measures for each user tend to be very close to the average. Thus the small values of standard deviations indicate that the metrics are also promising for each user individually.

The experiment shows that the execution time of the approach linearly increases as the database size increase, meaning that our approach have relatively good scale-up behavior since the increase of the number of learners in the database will lead to approximately the linear growth of the processing time, which is desirable in the processing of large databases.

5.4 Conclusion

In the first section of this chapter we demonstrate the effectiveness of our approach which exploited the strength of social aspect in folksonomies to let each member in the community benefit from the resources tagged by his other neighbors in the social networks based on resources recommendation. We have seen that it is very important to analyze the users profile in order to realize a personalized recommendation can be adapted to each modification in the users' interests. We have tested our approach on two baseline datasets where we have obtained promising results.

In the second section, we tested the efficacy of exploiting the strength of social aspect within medical field in order to let each one benefit from the resources tagged by his other neighbors in the social networks based on resources recommendation. We have tested our approach on a real world application for diabetes disease where we have obtained good results.

Our investigations in the field of collaborative E-learning have allowed us to make a substantial contribution in which we are interested to personalize resources retrieval according to learners' levels and specialties. We have tested the approach on a baseline dataset and we have obtained promising results.

Chapter 6: Conclusion

Chapter 6:
Conclusion

Information Retrieval is still a growing research area in computer science. Specifically, classic models of IR are about to evolve with the socialization of the web. In this context, the presented thesis investigated the social information retrieval in order to enhance and improve the classic recommender systems and information retrieval process within social technologies like folksonomies. Below, Section 6.1 presents a summary of our contributions and then Section 6.2 presents the possible future directions of our research work.

6.1 Contributions

This thesis is set in the research effort to bridge social web with information retrieval. In particular, we aimed at improving social information retrieval which is defined as the bridge that fills the gap between information retrieval and the web 2.0.

Social tagging which is the keystone of the social practices in web 2.0 has been highly developed in the last few years. Based on the motivation stated in chapter 1, our main research question was: how to improve recommendation and the retrieval effectiveness when searching personalized resources within social web applications?

In the presented thesis, we tried to resolve this research question by (1) analyzing users' profiles in social networks to help understanding tags' semantic during resource retrieval and recommendation, and (2) explicitly modeling the event dimension into resources ranking. Hence, the research questions we addressed were corresponding to two topics in social web: resources recommendation and ranking models within folksonomies. More specific research questions and solutions are presented below:

6.1.1 Resources recommendation in folksonomies

As it is motioned in earlier chapters, despite the strength of folksonomies, there are some problems hindering the growth of these systems: tag ambiguity, spelling variations and the lack of semantic links between tags.

Applying recommender systems within folksonomies recognized a real success in the last years, however due to the cited folksonomies problems it is clear that resource retrieval and so resource recommendation within folksonomies needs some improvements to increase the quality of the results obtained in these systems. The first research question we addressed was:

How to analyze user profiles within folksonomies to overcome tag ambiguity and spelling variation problems and thus improve the effectiveness of recommending personalized resources?

To answer this question, we proposed a new method to analyzing user profiles according to their tagging activity in order to improve resource recommendation within folksonomies. We based upon association rules which are a powerful method to discover interesting relationships among large datasets on the web. Focusing on association rules we found correlations between tags in a social network. Our aim was recommending resources annotated with tags suggested by association rules, in order to enrich user profiles. The effectiveness of the recommendation depends on the resolution of social tagging drawbacks. In our recommender process, we demonstrated how we can reduce tag ambiguity and spelling variations problems by taking into account social similarities calculated on folksonomies, in order to personalize resource recommendation. We surmounted also the lack of semantic links between tags during the recommendation process. We tested our approach on two baseline datasets where we obtained good results.

In order to get information about users' feedback on the recommender system, we proposed a real world application for diabetes disease. Diabetes affects millions of people in the world leading to substantial negative effects and expensive healthy penalties in our life. Recently with the emergence of social networks in the internet and their use in different field, we wanted to use this technology in clinical practice by showing a system based on giving doctors relevant medical resources can be annotated by them. Thus the second research question we addressed was:

How to help doctors discovering the best practices to patient diseases, diagnosis and treatments by analyzing doctors' profiles according to their daily tagging activity in order to personalize recommendation of medical resources?

We tried to answer this question by exploiting the above approach in medical field. We proposed a social application intended for all doctors with their different level of specialization and expertise. We argued that the automatic sharing of resources strengthens social links among users, and we exploited this idea to assemble doctors around a single web application to let each one of them benefit from the others' knowledge and expertise.

The usefulness of this application appeared in helping doctors to find the appropriate questions in the medical interview, helping them to make an appropriate diagnostic and also to propose a best treatment. The results carried out to get information about users' feedback on the recommender system are very promising.

6.1.2 Retrieval and Ranking Models

In many cases, when searching web resources, search results are displayed in chronological order where recently created resources are ranked higher than older ones. However, chronological ordering is not always effective. Therefore, a retrieval model should rank resources by the degree of relevance with respect to time. More precisely, documents must be ranked according to both social and temporal similarity.

An event-aware ranking model should take into account event detection, which captures the fact that the relevance of resources may change over time according to new events detection. Thus, the third research question we addressed is:

How to explicitly model the event dimension into resources ranking?

To answer this question, we tried to improving resources retrieval by exploiting the dependence between event detection and the high number of similar queries transmitted in the same period.

Folksonomies allow users to distribute and receive significant resources about real-world events. This content may appear in various forms, including status updates, photos, and videos, that can be created or posted before, during, and after an event. Furthermore, for known and planned events, structured information (e.g., title, time, location) might be available through event-aggregation social media sites. Such prior knowledge, however, is not available for unknown or spontaneous events (e.g., natural disasters). By automatically identifying the social media content related to either known or unknown events. It is clear that with the presence of a particular event, the number of similar queries in search engines increases considerably. We used a formula that can detect the presence of an event related to a given query since the increase of similar queries in a particular period.

In general, an event-aware ranking model gives scores to resources with respect to temporal feature. However, we wanted to study whether exploiting other features together with event detection can help improving the retrieval effectiveness in searching resources within social applications. In this case, we needed to find features used for capturing the similarity between an information need and personalization of retrieved resources, and combine such features with event dimension for relevance ranking. Thus, the last research question we addressed in this thesis was:

How to combine other features like social similarities with event detection in order to improve relevance ranking?

To achieve a good ranking, we sorted the retrieved resources according to some criterion so that the most relevant results appear early in the retrieved list displayed to the user. Two cases are observed: There is not an event detected during the search phase, in this case the retrieved resources are ranked only according to social similarities values. Otherwise, the proposed ranking function incorporated two features to be effective: the social aspect and the popularity of each resource during the same time period of event detection. A parameter is introduced to represent the importance one wants to give to the two types of features, i.e. social similarities or most popular resource in *n* time unites. In fact, depending on the context, one may want to give a higher importance to users' similarities. Another user may want to give more importance to special events that can be occurring, and so prefers the most popular resources.

The introduction of last feature in the ranking function is crucial, because the presence of an event can have a real influence on the popularity of resources and thus influenced the meaning of tags in the query even if those later are ambiguous.

6.2 Future work

Besides the contributions presented above, short term and long term perspectives are still to be investigated. In the context of the problems tackled in this PhD thesis (problems related to social web search and recommendation), we envision some perspectives related to each of our contributions as follows:

- In order to improve our approach, we aim at integrating additional automatic processing methods for the detection of semantic relations. This can reinforce the results we obtained.
- Evaluating our approach over other datasets, e.g. CiteULike, Last.fm, Bibsonomy, etc. Evaluating on different datasets has two main purposes: (i) Confirming and consolidating the performance and the results obtained on different datasets. (ii) Studying and illustrating the approach' behavior on other topologies of social networks in general and social tagging systems in particular.
- Evaluating our approach over semantic datasets. In this context, to improve association rules-based recommendation, it is interesting to propose an efficient frequent-pattern mining algorithm in semantic data. The proposed algorithm should handle all kinds of datasets and ontologies regardless of the dataset's domain. This should be an interesting direction for further study.
- Evaluating the approach on group recommender systems.

- Use the map-reduce framework in order to let the approach scale with very large databases.

Some long term perspectives are in the context of social recommendation. In particular, the topic that might interests our research team is to tackle the problem of diversity of information, i.e. in order not to annoy users with similar information. Indeed, the problem is a feeling that users have when they use Facebook in particular. Most of the time, when a recent information appears, all Facebook friends of a particular user begin to publish articles that deal with the same information, and the user is quickly overwhelmed by similar information. We believe that, at a given time, the recommender system should know that the user is already aware about this information and consequently it should be hidden.

Dealing with such issues is very interesting and motivating for our research team, even if we don't know how at the current moment. Do we have to deal with them by enhancing other collaborative filtering algorithms such as KNN or SVM for example? Or maybe we should propose new social recommendation algorithms? We believe that these problems should be deeply investigated.

Bibliography

Bibliography

[Abdelhaq et al., 2013] H. Abdelhaq, C. Sengstock, and M. Gertz. (2013) ‘EvenTweet: Online Localized Event Detection from Twitter’. In Proc. of the 39th International Conference on Very Large Data Bases (VLDB’13), Vol. 6, No. 12.

[Abel et al., 2008] F. Abel, N. Henze, and D. Krause. (2008) Ranking in folksonomy systems: can context help? In Proceedings of the 17th ACM conference on Information and knowledge management, CIKM ’08, pages 1429–1430, New York, NY, USA, 2008. ACM.

[Abiteboul et al., 2003] Abiteboul, S., Preda, M., & Cobena, G. (2003). Adaptive On-Line Page Importance Computation. Paper presented at the Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary.

[Adomavicius and Tuzhilin, 2005] G. Adomavicius and A. Tuzhilin, (2005) “Toward the next generation of recommender systems: a survey of the state-of-theart and possible extensions,” IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749.

[Adomavicius et al., 2011] G. Adomavicius, A. Tuzhilin (2011) Context-Aware Recommender Systems. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 217–253.

[Agraval et al., 1993] R. Agraval, T. Imielinski, and A. Swami. (1993) Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD Int. Conference on Management of Data, Washington, USA.

[Allan, 1996] J. Allan (1996) ‘Incremental relevance feedback for information filtering’. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland 270–278.

[Alonso and Shiells, 2013] O. Alonso and K. Shiells, (2013) ‘Timelines as Summaries of Popular Scheduled Events’. In Proc. of the International World Wide Web Conference (IW3C2).

[Alonso et al., 2007] O. Alonso, M. Gertz, and R. A. Baeza-Yates (2007). On the value of temporal information in information retrieval. SIGIR Forum, 41:35–41.

[Angeletou et al., 2007] S. Angeletou, M. Sabou, L. Specia, and E. Motta. (2007) Bridging the Gap Between Folksonomies and the Semantic Web: An Experience Report. In Proc. of ESWC workshop on Bridging the Gap between Semantic Web and Web.

[Anick, 2003] P. Anick (2003) 'Using terminological feedback for web search refinement: a log-based study'. In: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, New York, NY, USA, ACM Press 88–95.

[Antoniou et al., 2010] L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache (2010) The Semantic Web: Research and Applications: 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30 –June 3, 2010, Proceedings, Part I. Lecture Notes in Computer Science, vol. 6088, Springer-Verlag, Berlin Heidelberg New York.

[Baeza-Yates, B. Ribeiro-Neto, 1999] R. Baeza-Yates, B. Ribeiro-Neto, (1999). Modern information retrieval. New York: ACM Press.

[Baeza-Yates and Ribeiro-Neto, 2011] R.A. Baeza-Yates, B.A. Ribeiro-Neto. (2011) Modern Information Retrieval – the concepts and technology behind search, Second edition. Pearson Education Ltd., Harlow, England.

[Balakrishnan and Chopra, 2012] S. Balakrishnan, S. Chopra (2012) 'Collaborative ranking'. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM 2012, pp.143–152. ACM, New York.

[Bao et al., 2007] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. (2007) 'Optimizing web search using social annotations'. In Proceedings of the 16th international conference on World Wide Web, WWW '07, pages 501–510, New York, NY, USA, 2007. ACM.

[Becker et al., 2011] H. Becker, M. Naaman and L. Gravano. (2011) 'Beyond trending topics: Real-world event identification on Twitter'. In Proc of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM '11).

S. Beldjoudi, H. Seridi.les folksonomies entre la sémantique et l'effet communautaire. In Proc. Of JED, 2010.

[Beldjoudi et al., 2011a] S. Beldjoudi, H. Seridi and C. Faron-Zucker. (2001): Ambiguity in Tagging and the Community Effect in Researching Relevant Resources in Folksonomies. In Proc. of ESWC workshop User Profile Data on the Social Semantic Web.

[Beldjoudi et al., 2011b] S. Beldjoudi, H. Seridi and C. Faron-Zucker. (2011): Improving Tag-based Resource Recommendation with Association Rules on Folksonomies. In Proc. of ISWC workshop on Semantic Personalized Information Management: Retrieval and Recommendation.

[Beldjoudi et al., 2012a] S. Beldjoudi, H. Seridi and C. Faron-Zucker. (2012): Personalizing and Improving Tag-based Search in Folksonomies. In Proc. Of the 15th International Conference on Artificial Intelligence Methodology, Systems, Applications (AIMSA), Springer LNAI 7557, pp. 112–118.

[Beldjoudi et al., 2012b] S. Beldjoudi, H. Seridi and C. Faron-Zucker. (2012) The Social Semantic web between the meaning and the mining. In Proc. Of the COSI 2012 conference (Colloque sur l'Optimisation et les Systèmes d'Information).

[]Beldjoudi et al., 2012c] S. Beldjoudi, H. Seridi and C. Faron-Zucker. (2012) Let Tagging be more Interesting. In Proc. Of the IEEE Second International Workshop on Advanced Information Systems for Enterprises (IWAISE).

[Beldjoudi et al., 2014a] S. Beldjoudi, H. Seridi and C. Faron-Zucker. (2015) Analyzing Users' Profiles to Personalizing Resources Retrieval in Folksonomies, Accepted paper in the International Journal of Knowledge and Learning. (To appear).

[Beldjoudi et al., 2014b] S. Beldjoudi, H. Seridi and C. Faron-Zucker. (2015) Personalizing and Improving Resource Recommendation by Analyzing Users Preferences in Social Tagging Activities, Accepted paper in Computiong and Informatics Journal (Accepted paper).

[Bender et al., 2008] M. Bender, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. Xavier Parreira, R. Schenkel, and G. Weikum. (2008): Exploiting social relations for query expansion and result ranking. In ICDE Workshops, pages 501–506. IEEE Computer Society.

[Berberich et al., 2010] K. Berberich, S. J. Bedathur, O. Alonso, and G. Weikum. (2010) A language modeling approach for temporal information needs. In Proceedings of the 32nd European Conference on IR Research on Advances in Information Retrieval, ECIR ’10, pages 13–25.

[Bharat et al., 1998] K. Bharat, T. Kamba, M. Albers (1998) ‘Personalized, interactive news on the web’. Multimedia Syst. 6(5) 349–358.

[Bischoff et al., 2008] K. Bischoff, C.S. Firan, W. Nejdl, R. Paiu (2008) Can all tags be used for search? In: CIKM.

[Blei et al., 2003] D.M. Blei, A.Y. Ng, M.I Jordan (2003) ‘Latent dirichlet allocation’. J. Mach. Learn. Res. 3, 993–1022.

[Boff and Reategui, 2012] E. Boff, E. Reategui (2012) ‘Mining Social-Affective Data to Recommend Student Tutors’. In Proceedings of the 13th Ibero-American Conference on AI, Cartagena de Indias, Colombia, November 13-16, 2012, PP 672-681.

[Bouadjenek et al., 2013] M.R. Bouadjenek, H. Hacid, and M. Bouzeghoub (2013) SoPRa: A New Social Personalized Ranking Function for Improving Web Search. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’13, pages 861–864, New York, NY, USA, 2013. ACM.

[Boughareb and Farah, 2012] D. Boughareb, N. Farah (2012): Contextual Modelling of the User Browsing Behaviour to Identify the User’s Information Need. In Proceedings of the Second International Conference on Innovative Computing Technology (INTECH), 2012 (IEEE) pp 247 – 252.

[Breese et al., 1998] J.S. Breese, D. Heckerman, and C. Kadie. (1998): Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, July 1998.

[Broder, 2002] Broder, A. (2002). A taxonomy of web search. Paper presented at the Proceedings of the ACM SIGIR Forum.

[Buckland, 1999] M. Buckland. (1999): “Vocabulary as a Central Concept in Library and Information Science.” Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the Third International Conference on Conceptions of Library and Information Science (CoLIS3, Dubrovnik, Croatia, 23-26 May 1999. Ed. by T. Arpanac et al. Zagreb: Lokve, pp 3-12.

[Buffa et al., 2008] M. Buffa, F. Gandon, G. Ereteo, P. Sander, and C. Faron. (2008): SweetWiki: A semantic Wiki. Journal of Web Semantics.

[Burke et al., 2005] R. Burke, B. Mobasher, R. Bhaumik, and C. Williams. (2005): Segment-based injection attacks against collaborative filtering recommender systems. In ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining, pages 577–580, Washington, DC, USA. IEEE Computer Society.

[Cantador et al., 2011] I. Cantador, P. Castells (2011) Extracting Multilayered Communities of Interest from Semantic User Profiles: Application to Group Modeling and Hybrid Recommendations. In: Computers in Human Behavior. Elsevier.

[Carmel et al., 2009] D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'el, I. Ronen, E. Uziel, S. Yogev, and S. Chernov. (2009): Personalized social search based on the user's social network. In Proceedings of the 18th ACM conference on Information and knowledge management, CIKM '09, pages 1227–1236, New York, NY, USA. ACM.

[Cattuto et al., 2008] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. (2008): Semantic grounding of tag relatedness in social bookmarking systems. In ISWC '08: Proceedings of the 7th International Conference on The Semantic Web, p. 615–631, Berlin, Heidelberg:Springer-Verlag.

[Chan and Jin, 2006] S.M. Chan, Q. Jin (2006) 'Collaboratively Shared Information Retrieval Model for e-Learning'. In Proceedings of the 5th International Conference, Penang, Malaysia, July 19-21.

[Chavarriaga et al., 2014] O. Chavarriaga, B. Florian-Gaviria, and O. Solarte (2014) 'A Recommender System for Students Based on Social Knowledge and Assessment Data of Competences'. In Proceedings of the 9th European Conference on Technology Enhanced Learning, EC-TEL 2014, Graz, Austria, September 16-19.

[Chomutare, 2014] T. Chomutare (2014) 'Patient Similarity Using Network Structure Properties in Online Communities'. In Proceedings of the International Conference of

Biomedical and Health Informatics (BHI), 2014 IEEE-EBMS PP 809-812.

[Cremonesi et al., 2010] P. Cremonesi, Y. Koren, R. Turrin (2010) 'Performance of recommender algorithms on top-n recommendation tasks'. In: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys 2010, pp. 39–46. ACM, New York.

[Dakka et al., 2008] W. Dakka, L. Gravano, and P. G. Ipeirotis (2008). Answering general time sensitive queries. In Proceedings of the 17th ACM conference on Information and knowledge management, CIKM '08, pages 1437–1438.

[De Meo et al., 2010] P. De Meo, G. Quattrone, and D. Ursino. (2010): A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. User Modeling and User-Adapted Interaction, 20(1).

[Dennis et al., 2003] Dennis, F., Mark, M., & Marc, N. (2003). On the Evolution of Clusters of Near-Duplicate Web Pages. Paper presented at the Proceedings of the First Conference on Latin American Web Congress.

[Dieberger et al., 2000] A. Dieberger, P. Dourish, K. Hook, P. Resnick, A. Wexelblat (2000) 'Social navigation: techniques for building more usable systems'. Interactions **7**(6) 36–45.

[Ding et al., 2008] X. Ding, B. Liu, P.S. Yu (2008) A holistic lexicon-based approach to opinion mining. In: Proceedings of the conference on Web search and Web data mining (WSDM'08). ACM, Palo Alto, pp 231–24050.

[Doctorow et al., 2002] C. Doctorow, F. Dornfest, J. Johnson, S. Powers. (2002): Essential Blogging. O'Reilly.

[Donciu et al., 2011] M. Donciu, M. Ionita, M. Dascalu, S. Trausan-Matu (2011) 'The Runner - Recommender system of workout and nutrition for runners'. In Proceedings of the 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (IEEE).

[Dong et al., 2010] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. (2010) 'Time is of the essence: improving recency ranking using twitter data'. In Proceedings of the 19th international conference onWorld wide web,WWW'10, pages 331–340, New York, NY, USA, ACM.

[Ebersbach et al., 2006] A. Ebersbach, M. Glaser, R. Heigl. (2006): Wiki: Web Collaboration. Springer-Verlag: Germany.

[Euzenat and Shvaiko, 2007] J. Euzenat, P. Shvaiko (2007) ‘Ontology Matching’. Berlin, Heidelberg: Springer.

[Filho et al., 2010] F.M.F Filho, G.M. Olson, P.L. de Geus (2010) Kolline: a task-oriented system for collaborative information seeking. SIGDOC 2010: 89-94.

[Firan et al., 2007] C. Firan, W. Nejdl, R. Paiu (2007) The benefit of using tagbased profiles. In: Proceedings of the 2007 Latin American web conference (LA-WEB 2007). Santiago de Chile, Chile, pp 32–41.

[Freyne and Smyth, 2004] J. Freyne, B. Smyth (2004) ‘An experiment in social search’. In Bra, P.D., Nejdl,W., eds.: Adaptive Hypermedia and Adaptive Web-Based Systems, Third International Conference, AH 2004, Eindhoven, The Netherlands, August 23-26, 2004, Proceedings. Volume 3137 of Lecture Notes in Computer Science., Springer 95–103.

[Furnas et al., 1987] G.W. Furnas, T.K. Landauer, L.M. Gomez, S.T. Dumais (1987) ‘The vocabulary problem in human-system communication’. Commun. ACM 30(11) 964–971.

[Gamon et al., 2005] M. Gamon, A. Aue, S. Corston-Oliver, E. Ringger (2005) Pulse: mining customer opinions from free text. pp 121–132.

[Gartrell et al., 2010] M. Gartrell, X. Xing, Q. Lv, A. Beach, R. Han, S. Mishra, K. Seada (2010) Enhancing Group Recommendation by Incorporating Social Relationship Interactions. In: Proceedings of the 16th ACM International Conference on Supporting Group Work (GROUP 2010), pp. 97–106.

[Gemmell et al., 2009] J. Gemmell, T. Schimoler, M. Ramezani, and B. Mobasher (2009). Adapting k-nearest neighbor for tag recommendation in folksonomies. In Proc. of 7th Workshop on Intelligent Techniques for Web Personalization \& Recommender Systems (ITWP 09), Pasadena, California, USA, in conjunction with IJCAI.

[Geyer et al., 2010] W. Freyne, J. Mobasher, B. Anand, S.S. Dugan (2010): 2nd Workshop on Recommender Systems and the Social Web. In: Proceedings of the 4th ACM Conference on Recommender Systems (RecSys 2010), pp. 379–380.

[Gnasa, 2006] Gnasa, M. (2006). Congenial Web Search - A Conceptual Framework for Personalized, Collaborative, and Social Peer-to-Peer Retrieval. Dissertation, Rheinischen Friedrich-Wilhelms-Universität Bonn, Bonn.

[Goh and Foo, 2007] H.L.D Goh, S. and Foo (2007) Social Information Retrieval Systems: emerging technologies and applications for searching the web effectively, Information Science Reference, an imprint of IGI Global.

[Golder and Huberman, 2006] S. Golder, B.A. Huberman (2006) Usage patterns of collaborative tagging systems. Journal of Information Science, 32(2):198--208.

[Goldberg et al., 1992] D. Goldberg, D. Nichols, B.M. Oki, D. Terry (1992) 'Using collaborative filtering to weave an information tapestry'. Commun. ACM 35(12) 61–70.

[Gong and Sun., 2011] J. Gong, S. Sun (2011) 'Individual Doctor Recommendation Model on Medical Social Network'. In Proceedings of the 7th International Conference, ADMA 2011, Beijing, China, December 17-19, 2011, PP 69-81.

[Gong et al., 2014] J. Gong, C. Pang, L. Wang, L. Zhang, W. Huang, S. Sun (2014) 'Doctor Recommendation via Random Walk with Restart in Mobile Medical Social Networks'. H.-Y. Huang et al. (Eds.): SMP 2014, CCIS 489, pp. 198–205, 2014.

[Gruber, 1993] T. Gruber, (1993): A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2), pp. 199-220.

[Hamed et al., 2012] A.A. Hamed. R. Roose, M. Branicki, A. Rubin, MD (2012) 'T-Recs: Time-aware Twitter-based Drug Recommender System'. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[Harpale and Yang, 2008] A.S. Harpale and Y. Yang. (2008): Personalized active learning for collaborative filtering. In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 91–98, New York, NY, USA, ACM.

[Haveliwala, 2002] Haveliwala, T. H. (2002). Topic-sensitive PageRank. Paper presented at the Proceedings of the 11th international conference on World Wide Web, Honolulu, Hawaii, USA.

[Hill et al., 1995] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, 1995: "Recommending and Evaluating Choices in a Virtual Community of Use," Proc. Conf. Human Factors in Computing Systems.

[Hoffman, 2008] T. Hoffman (2008) Online reputation management is hot—but is it ethical? ComputerWorld.

[Hotho et al., 2006] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. (2006): Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, editors, The SemanticWeb: Research and Applications, pages 411–426.

[Huang et al., 2011] C.L Huang, H.Y Chien, M Conyette, (2011): Folksonomy-based Recommender Systems with User's Recent Preferences, World Academy of Science, Engineering and Technology 78.

[Jansen et al., 2010] B.J. Jansen, A. Chowdury, and G. Cook. (2010) 'The ubiquitous and increasingly significant status message. Interactions', 17(3):15–17.

[Jäschke et al., 2007] R. Jaschke, L.B. Marinho, Hotho A., L. Schmidt-Thieme, and G. Stumme. (2007): Tag recommendations in folksonomies. In Proc. of 11th Eur. Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Warsaw, Poland, volume 4702 of LNCS. Springer.

[Jin and Si, 2004] R. Jin and L. Si. (2004): A bayesian approach toward active learning for collaborative filtering. In UAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence, pages 278–285, Arlington, Virginia, United States. AUAI Press.

[Joachims, 2006] T. Joachims (2002). Unbiased evaluation of retrieval quality using clickthrough data: Cornell University, Department of Computer Science.

[Joachims et al., 2005] T. Joachims, L. Granka, B. Pan, H. Hembrooke, G. Gay (2005). Accurately interpreting clickthrough data as implicit feedback. Paper presented at the Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05), Salvador, Brazil.

[Jong et al., 2005] F. de Jong, H. Rode, and D. Hiemstra. (2005) Temporal language models for the disclosure of historical text. In Humanities, computers and cultural heritage: Proceedings of the 16th International Conference of the Association for History and Computing (AHC 2005), pages 161–168.

[Joshi and Belsare, 2006] M. Joshi, N. Belsare (2006) Blogharvest: Blog mining and search framework. In: International conference on management of data. Computer Society of India, Delhi, pp 226–229.

[Kamvar et al., 2003] S.D. Kamvar, T.H, Haveliwala, C.D. Manning, G.H. Golub, (2003). Extrapolation methods for accelerating PageRank computations. Paper presented at the Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary.

[Kaplan, 2001] B. Kaplan (2001) 'Evaluating informatics applications—clinical decision support systems literature review': International Journal of Medical Informatics 64 (2001) 15–37.

[Kaplan and Haenlein, 2010] A. Kaplan, M. Haenlein, (2010): "Users of the world, unite! The challenges and opportunities of Social Media". Business Horizons 53(1): 59–68.

[Kleinberg, 1999] Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5), 604-632.

[Köck and Paramythis, 2011] M. Köck, A. Paramythis (2011) 'Activity sequence modelling and dynamic clustering for personalized e-learning'. User Modeling and User-Adapted Interaction April 2011, Volume 21, Issue 1-2, pp 51-97.

[Koenemann and Belkin, 1996] J. Koenemann, N.J. Belkin (1996) 'A case for interaction: a study of interactive information retrieval behavior and effectiveness'. In: CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM Press 205–212.

[Koerner et al., 2010] C. Koerner, D. Benz, M. Strohamaier, A. Hotho and G. Stumme. (2010): Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In Proceedings of the 19th International World Wide Web Conference (WWW), Raleigh, NC, USA: ACM.

[Koren, 2010] Y. Koren (2010) ‘Factor in the neighbors: Scalable and accurate collaborative filtering’. ACM Trans. Knowl. Discov. Data 4(1), 1:1–1:24.

[Koren and Sill, 2011] Y. Koren, J. Sill (2011) ‘Ordrec: an ordinal model for predicting personalized item rating distributions’. In: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys 2011, pp. 117–124. ACM, New York.

[Krause et al., 2008] B. Krause, A. Hotho, G. Stumme (2008) A comparison of social bookmarking with traditional search. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 101–113. Springer, Heidelberg.

[Kulkarni et al., 2011] A. Kulkarni, J. Teevan, K. M. Svore, and S. T. Dumais. (2011) Understanding temporal query dynamics. In Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM ’11, pages 167–176.

[Lam and Riedl, 2004] S.K. Lam and J. Riedl. (2004) Shilling recommender systems for fun and profit. In WWW ’04: Proceedings of the 13th international conference on World Wide Web, pages 393–402, New York, NY, USA, ACM.

[Lawrence and Giles] S. Lawrence, C.L. Giles (1998) ‘Context and page analysis for improved web search’. IEEE Internet Computing 2(4) 38–46.

[Lewandowski, 2005] D. Lewandowski (2005). Web searching, search engines and Information Retrieval. Information Science and Use, 25(3), 137-147.

[Li and Iribe, 2012] K. Li, Y. Iribe (2012) ‘Supporting Continued Communication with Social Networking Service in e-Learning’. Intelligent Interactive Multimedia: Systems and Services Smart Innovation, Systems and Technologies Volume 14, 2012, pp 569-577.

[Li and Zaman, 2014] J. Li, N. Zaman (2014) ‘Personalized Healthcare Recommender Based on Social Media’. In Proceedings of the IEEE 28th International Conference on Advanced Information Networking and Applications.

[Lim and Husain, 2010] T.P. Lim, W. Husain (2010) ‘Integrating knowledge-based system in Wellness Community Portal’. In Proceedings of the International Conference on Science and Social Research (CSSR 2010), December 5 - 7, 2010, Kuala Lumpur, Malaysia.

[Limpens et al., 2010] F. Limpens, F. Gandon, and M. Buffa. (2010): Collaborative semantic structuring of folksonomies. In Proc. of IEEE/WIC/ACM Int. Conference on Web Intelligence (WI).

[Liu, 2009] T.-Y. Liu. (2009) Learning to rank for information retrieval. Found. Trends Inf. Retr., 3(3):225–331.

[Liu, 2011] T.Y. Liu (2011) 'Learning to rank for information retrieval', Berlin, German, vol. 3(3), pp. 225–331. Springer.

[Liu and Yang, 2008] N.N. Liu, Q. Yang (2008) 'Eigenrank: a ranking-oriented approach to collaborative filtering'. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 83–90. ACM, New York.

[Liu et al., 2004] Liu, F., Yu, C., & Meng, W. (2004). Personalized Web search for improving retrieval effectiveness. IEEE Transactions on Knowledge and Data Engineering, 16(1), 28- 40.

[Lopez-Nores et al., 1011] M. Lopez-Nores, Y. Blanco-Fernandez, J. J. Pazos-Arias, J. Garcia-Duque, and M.I. Martin-Vicente (2011) 'Enhancing Recommender Systems with Access to Electronic Health Records and Groups of Interest in Social Networks'. In Proceedings of the Seventh International Conference on Signal Image Technology & Internet-Based Systems (IEEE).

[Manning et al., 2009] C. D Manning, P. Raghavan, and H. Schutze. (2009): An introduction to information retrieval. Cambridge University Press.

[Massa and Bhattacharjee, 2004] P.Massa and B. Bhattacharjee (2004) Using trust in recommender systems: an experimental analysis. Proceedings of 2nd International Conference on Trust Managment, Oxford, England.

[Mathes, 2004] A. Mathes (2004) 'Folksonomies - Cooperative Classification and Communication Through Shared Metadata'. Rapport interne, GSLIS, Univ. Illinois Urbana-Champaign.

[Melville et al., 2002] P. Melville, R.J. Mooney, and R. Nagarajan. (2002): Content-boosted collaborative filtering for improved recommendations. In Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-02), pages 187–192, Edmonton, Alberta.

[Metzler et al., 2009] D. Metzler, R. Jones, F. Peng, and R. Zhang. Improving search relevance for implicitly temporal queries. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’09, pages 700–701, 2009.

[Micarelli and Sciarrone, 2004] A. Micarelli, F. Sciarrone (2004) ‘Anatomy and empirical evaluation of an adaptive web-based information filtering system’. User Modeling and User-Adapted Interaction 14(2-3) 159–200.

[Mika, 2005] P. Mika. (2005): Ontologies are us: A unified model of social networks and semantics. In Proc. of 4th Int. Semantic Web Conference (ISWC 2005), Galway, Ireland, volume 3729 of LNCS. Springer.

[Mooney and Roy, 2000] R.J. Mooney and L. Roy. (2000): Content-based book recommending using learning for text categorization. In Proceedings of the Fifth ACM Conference on Digital Libraries, pages 195–204, San Antonio, TX.

[Morris, 2007] MR. Morris. (2007) ‘Collaborating alone and together: Investigating persistent and multi-user web search activities’. Technical report, Microsoft Research.

[Morris, 2008] M.R. Morris. (2008) ‘A survey of collaborative web search practices’. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’08, pages 1657–1660, New York, NY, USA, ACM.

[Morris and Horvitz, 2007] M.R. Morris and E. Horvitz. (2007) ‘Searchtogether: an interface for collaborative web search’. In Proceedings of the 20th annual ACM symposium on User interface software and technology, UIST ’07, pages 3–12, New York, NY, USA, ACM.

[Moutachaouik et al., 2012] H. Moutachaouik, H. Douzi, A. Marzak, H. Behja, and B. Ouhbi (2012) ‘Plugin of Recommendation Based on a Hybrid Method for the Ranking of Documents in the E-Learning Platforms‘. In Proceedings of the 5th International Conference, ICISP 2012, Agadir, Morocco, June 28-30.

[Mutschke and Mayr, 2015] P. Mutschke, P. Mayr (2015) ‘Science models for search: a study on combining scholarly information retrieval and scientometrics’. In Journal Scientometrics Volume 102, Issue 3 , pp 2323-2345.

[Noll and Meinel, 2007] M.G. Noll, C. Meinel (2007) ‘Web search personalization via social bookmarking and tagging’. In: Aberer, K., et al. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 367–380. Springer, Heidelberg.

[Nørvåg, 2004] K. Nørvåg. Supporting temporal text-containment queries in temporal document databases. Journal of Data & Knowledge Engineering, 49(1):105–125, April 2004.

[Olston and Chi,, 2003] C. Olston, E.H. Chi (2003) ‘ScentTrails: Integrating browsing and searching on the web’. ACM Transactions on Computer-Human Interaction 10(3)177–197.

[Palazuelos et al., 2013] C. Palazuelos, D. Garcia-Saiz, and M. Zorrilla (2013) ‘Social Network Analysis and Data Mining: An Application to the E-Learning Context’. In Proceedings of the 5th International Conference, ICCCI 2013, Craiova, Romania, September 11-13.

[Pan et al., 2010] J.Z. Pan, S. Taylor, and E. Thomas. (2010): Reducing ambiguity in tagging systems with folksonomy search expansion. In Proc. of 6th Eur. Semantic Web Conference (ESWC), Heraklion, Greece, volume 5554 of LNCS. Springer.

[Pasca, 2007] Pasca, M. (2007). Organizing and Searching the World Wide Web of Facts - Step Two: Harnessing the Wisdom of the Crowds. Paper presented at the Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada.

[Passant, 2009] A. Passant (2009) ‘Technologies du Web Sémantique pour l’Entreprise 2.0’. PhD thesis, Université Paris IV - Sorbonne.

[Pau and Morris, 2009] S.A. Paul, and M.R. Morris (2009). CoSense: Enhancing Sensemaking for Collaborative Web Search. Proceedings of the Conference on Human Factors in Computing Systems (CHI 2009), Boston, MA.

[Pazzani and Billsus, 1997] M.J. Pazzani and D. Billsus. (1997): Learning and revising user profiles: The identification of interesting web sites. Machine Learning, 27(3):313–331.

[Pitkow et al., 2002] J. Pitkow, H. Sch¨utze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, T. Breuel (2002) 'Personalized search'. Commun. ACM 45(9) 50–55.

[Psallidas et al., 2013] F. Psallidas, H. Becker, M. Naaman, and L. Gravano (2013) Effective Event Identification in Social Media, in IEEE Data Engineering Bulletin -- Special Issue of Social Media Analytics (IEEE DEB 2013).

[Rendle et al., 2009] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme (2009) 'Bpr: Bayesian personalized ranking from implicit feedback'. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2009, Arlington, Virginia, United States, pp. 452–461. AUAI Press.

[Resnick et al., 1994] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl (1994) 'Grouplens: an open architecture for collaborative filtering of netnews'. In: CSCW '94: Proceedings of the 1994 ACM conference on Computer supported cooperative work, New York, NY, USA, ACM Press 175–186.

[Rijsbergen, 1979] C.J.V. Rijsbergen (1979) 'Information Retrieval'. Butterworth-Heinemann, Newton, MA, USA.

[Rivero-Rodríguez et al., 2013] A. Rivero-Rodríguez, S. Th. Konstantinidis, C.L. Sanchez-Bocanegra, L. Fernández-Luque (2013) 'A Health Information Recommender System: enriching YouTube Health Videos with Medline Plus Information by the use of SnomedCT terms'. In Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (*CBMS 2013*) Porto, Portugal.

[Robertson and Jones, 1976] S. E. Robertson and K. S. Jones (1976) Relevance weighting of search terms. Journal of the American Society for Information Science, 27(3):129–146.

[Sakaki et al., 2010] T. Sakaki, M. Okazaki, and Y. Matsuo. (2010) 'Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development'. IEEE Trans. Knowl. Data Eng. 25(4): 919-931.

[Salton and McGill, 1983] G. Salton, M. McGill (1983) An Introduction to modern information retrieval. Mc-Graw-Hill, New York.

[Schein et al., 2002] A. Schein, A. Popescul, L.H. Ungar, D.M. Pennock (2002) Methods and metrics for cold-start recommendations. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information, Tampere, Finland.

[Schmitz et al., 2006] C. Schmitz, A. Hotho, R. Jaschkee, and G. Stumme. (2006): Mining association rules in folksonomies. In Proc. of IFCS 2006 Conference: Data Science and Classification, Ljubljana, Slovenia. Springer.

[Shardanand and Maes, 1995] U. Shardanand and P. Maes, (1995): "Social Information Filtering: Algorithms for Automating 'Word of Mouth'," Proc. Conf. Human Factors in Computing Systems.

[Shi et al., 2010] Y. Shi, M. Larson, A. Hanjalic (2010) 'List-wise learning to rank with matrix factorization for collaborative filtering'. In: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys 2010, pp. 269–272. ACM, New York.

[Shi et al., 2012] Y. Shi, A. Karatzoglou, L. Baltrunas, M. Larson, N. Oliver, A. Hanjalic (2012) 'Climf: learning to maximize reciprocal rank with collaborative less-is-more filtering'. In: Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys 2012, pp. 139–146. ACM, New York.

[Shi et al., 2012] Y. Shi, A. Karatzoglou, L. Baltrunas, M. Larson, A. Hanjalic, N. Oliver (2012) 'Tfmap: optimizing map for top-n context-aware recommendation'. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2012, pp. 155–164. ACM, New York.

[Shi et al., 2013] L. Shi, D. Al Qudah, A. Qaffas, and A.I. Cristea (2013) 'Topolor: A Social Personalized Adaptive E-Learning System'. In Proceedings of the 21th International Conference, UMAP 2013, Rome, Italy, June 10-14, PP 338-340.

[Shi et al., 2014] L. Shi, A.I. Cristea, S. Hadzidedic, and N. Dervishalidovic (2014) 'Contextual Gamification of Social Interaction –Towards Increasing Motivation in Social E-learning'. In Proceedings of the 13th International Conference, Tallinn, Estonia, August 14-17.

[Silverstein et al., 1999] Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a very large web search engine query log. SIGIR Forum 33 (1999) 6–12.

[Song and Marsh, 2012] I. Song, N.V. Marsh (2012) 'Anonymous Indexing of Health Conditions for a Similarity Measure'. In Proceedings of the IEEE Transactions on Information Technology In Biomedicine, Vol. 16, No. 4.

[Song et al., 2011] I. Song, D. Dillon, T.J. Goh, and M. Sung (2011) 'A Health Social Network Recommender System'. In Proceedings of the 14th International Conference, PRIMA 2011, Wollongong, Australia, November 16-18, 2011. PP 361-372.

[Specia and Motta, 2007] L. Specia and E. Motta. (2007): Integrating folksonomies with the semantic web. In Proc. of 4th Eur. Semantic Web Conference (ESWC 2007), Innsbruck, Austria, volume 4519 of Lecture Notes in Computer Science. Springer.

[Speretta and Gauch, 2005] M. Speretta, S. Gauch (2005) 'Personalized search based on user search histories'. In: Web Intelligence (WI2005), France, IEEE Computer Society 622–628 http://dx.doi.org/10.1109/WI.2005.114.

[Spink and Jansen, 2004] A. Spink, B.J. Jansen (2004) 'A study of web search trends'. Webology 1(2) http://www.webology.ir/2004/v1n2/a4.html.

[Spink et al., 2000] A. Spink, B.J Jansen, H.C. Ozmultu (2000) 'Use of query reformulation and relevance feedback by excite users'. Internet Research: Electronic Networking Applications and Policy 10(4) 317–328 http://citeseer.ist.psu.edu/spink00use.html.

[Stewart et al., 2007] A. Stewart, L. Chen, R. Paiu, W. Nejdl (2007) Discovering information diffusion paths from blogosphere for online advertising. In: ADKDD '07: Proceedings of the 1st international workshop on data mining and audience intelligence for advertising. ACM, New York, pp 46–54.

[Stock, 2007] W.G. Stock (2007). Information Retrieval Informationen suchen und finden: Oldenbourg.

[Su et al., 2008] X. Su, T.M. Khoshgoftaar, X. Zhu, and R. Greiner. Imputation-boosted collaborative filtering using machine learning classifiers. In SAC '08: Proceedings of the 2008 ACM symposium on Applied computing, pages 949–950, New York, NY, USA, 2008. ACM.

[Szomszor et al., 2008] M. Szomszor, I. Cantador, H. Alani (2008) Correlating User Profiles from Multiple Folksonomies. In: Proceedings of the 19th ACM Conference on Hypertext and Hypermedia (Hypertext 2008), pp. 33–42.

[Szomszor et al., 2010] M. Szomszor, C. Cattuto, W. Van den Broeck, A. Barrat, H. Alani (2010) Semantics, Sensors, and the Social Web: The Live Social Semantics Experiments. In: Aroyo,

[Takahashi and Kitagawa, 2008] T. Takahashi and H. Kitagawa. (2008): S-bits: Social-bookmarking induced topic search. In Proceedings of the 2008 The Ninth International Conference on Web-Age Information Management, WAIM '08, pages 25–30, Washington, DC, USA. IEEE Computer Society.

[Takahashi and Kitagawa, 2009] T. Takahashi and H. Kitagawa. (2009): A ranking method for web search using social bookmarks. In Proceedings of the 14th International Conference on Database Systems for Advanced Applications, DASFAA '09, pages 585–589, Berlin, Heidelberg, Springer-Verlag.

[Tan et al., 2006] Tan, B., Shen, X., & Zhai, C. (2006). Mining long-term search history to improve search accuracy. Paper presented at the Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA.

[Tatemura, 2000] J. Tatemura (2000) Virtual reviewers for collaborative exploration of movie reviews. In: Proceedings of intelligent user interfaces (IUI). pp 272–275.

[Teevan et al., 2004] J. Teevan, C. Alvarado, M.S. Ackerman, D.R. Karger (2004) 'The perfect search engine is not enough: a study of orienteering behavior in directed search'. In: CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM Press 415–422.

[Teevan et al., 2005] J. Teevan, S.T. Dumais, E. Horvitz (2005) 'Personalizing search via automated analysis of interests and activities'. In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM Press 449–456.

[Teevan et al., 2011] J. Teevan, D. Ramage, and M.R. Morris. (2011) ‘#twittersearch: a comparison of microblog search and web search’. In Proceedings of the fourth ACM international conference on Web search and data mining, WSDM ’11, pages 35–44, New York, NY, USA.

[Thong and Son, 2015] N.T. Thong, L.H. Son (2015) ‘HIFCF: An effective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis’. Expert Systems with Applications 42 (2015) 3682–3701.

[Tomlin, 2003] Tomlin, J. A. (2003). A new paradigm for ranking pages on the world wide web. Paper presented at the Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary.

[Torniai et al., 2008] C. Torniai, J. Jovanović, D. Gašević, S. Bateman, and M. Hatala. (2008) ‘E-learning meets the Social Semantic Web’. In Proc. of Eighth IEEE International Conference on Advanced Learning Technologies.

[Trant, 2009] J Trant (2009) Studying social tagging and folksonomy: A review and framework . Journal of Digital Information (JoDI)

[Tso-Sutter et al., 2008] K.H.L. Tso-Sutter, L.B. Marinho, L. Schmidt-Thieme, (2008): Tag-aware recommender systems by fusion of collaborative filtering algorithms. In Proc. of the ACM Symposium on Applied Computing (SAC 2008), pp. 1995–1999. ACM Press, Fortaleza.

[Vallet et al., 2010] D. Vallet, I. Cantador and J.M. Jose. (2010): Personalizing web search with folksonomy-based user and document profiles. In Proceedings of the 32nd European conference on Advances in Information Retrieval, ECIR’2010, pages 420–431, Berlin, Heidelberg, Springer-Verlag.

[Van damme et al., 2007] C. Van damme, M. Hepp, K. Siorpaes (2007) ‘Folksontology: An integrated approach for turning folksonomies into ontologies’. In Bridging the Gep between Semantic Web and Web 2.0 (SemNet 2007), p. 57–70.

[Vander Wal, 2005] T. Vander wal (2005). "Off the Top: Folksonomy Entries." Visited April 5, 2015. See also: Smith, Gene. "Atomiq: Folksonomy: social classification".

[Vander Wal, 2007] T. Vander Wal, (2007): Folksonomy Coinage and Definition. Vanderwal.net.

[Versin et al., 2013] B. Versin, A. Klasnja-Milicevic, M. Ivanovic and Z. Budimac. (2013): Applying recommender systems and adaptive hypermedia for E-learning personalization. Journal of Computing and Informatics, Vol. 32, 629-659.

[Volkovs and Zemel, 2013] M.N. Volkovs, R.S. Zemel (2013) 'Collaborative ranking with 17 parameters'. In: Proceedings of the Twenty-sixth Annual Conference on Neural Information Processing Systems, NIPS 2013, Lake Tahoe, Nevada, United States, December 3-6, MIT Press.

[Wærn, 2004] A. Wærn (2004) 'User involvement in automatic filtering: An experimental study'. User Modeling and User-Adapted Interaction 14(2-3) 201–237.

[Wang and Jin, 2010] Q. Wang and H. Jin. (2010): Exploring online social activities for adaptive search personalization. In Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM'10, pages 999–1008, New York, NY, USA, ACM.

[Wang et al., 2004] Wang, Y., & DeWitt, D. J. (2004). Computing pagerank in a distributed internet search system. Paper presented at the Proceedings of the Thirtie13th international conference on Very large data bases, Toronto, Canada.

[Wang et al., 2010] Y. Wang, M. Zhu, L. Qu, M. Spaniol, and G. Weikum. Timely YAGO (2010) harvesting, querying, and visualizing temporal knowledge from Wikipedia. In Proceedings of the 13th International Conference on Extending Database Technology, EDBT '10, pages 697–700.

[Wang et al., 2011] G. Wang, H. Wang, X. Tao, and J. Zhang (2011) 'Positive Influence Dominating Set in E-Learning Social Networks'. In Proceedings of the 10th International Conference, Hong Kong, China, December 8-10.

[Wang et al., 2012] Y. Wang, H. Sundaram and L. Xie, (2012) 'Social Event Detection with Interaction Graph Modeling' In Proc. of ACM International Conference on Multimedia, Association for Computing Machinery Inc (ACM), Nara Japan, pp. 1-4.

[Webb et al., 2001] G.I. Webb, M. Pazzani, D. Billsus (2001) 'Machine learning for user modeling'. User Modeling and User-Adapted Interaction 11(1-2) 19–29.

[Weimer et al., 2009] M. Weimer, A. Karatzoglou, M. Bruch (2009) 'Maximum margin matrix factorization for code recommendation'. In: Proceedings of the Third ACM Conference on Recommender Systems, RecSys 2009, pp. 309–312. ACM, New York.

[Westerski et al., 2006] A. Westerski, S.R. Kruk, K. Samp, T. Woroniecki, F. Czaja, and C. O'Nuallain. (2006) 'E-learning based on the social semantic information sources.' In Proc. of LACLO'2006.

[Winoto et al., 2008] P. Winoto, T. Ya Tang (2008) If You Like the Devil Wears Prada the Book, Will You also Enjoy the Devil Wears Prada the Movie? A Study of Cross-Domain Recommendations. New Generation Computing 26(3), 209–225.

[Wong et al., 2013] K. Wong, R. Kwan, F.L. Wang, and L. Luk (2013) 'Students' Experience and Perception on E-Learning Using Social Networking'. In Proceedings of 6th International Conference, ICHL 2013, Toronto, ON, Canada, August 12-14, pp 269-279.

[Wu and Zhou, 2011] C. Wu and B. Zhou. (2011): Tags are related: Measurement of semantic relatedness based on folksonomy network. Journal of Computing and Informatics, Vol. 30, 165-188.

[Xu et al., 2006] Z. Xu, Y. Fu, J. Mao, D. Su (2006) Towards the semantic web: collaborative tag suggestions. In: Proceedings of the 15th international WWW conference. Collaborative Web Tagging Workshop.

[Xu et al., 2008] S. Xu, S. Bao, B. Fei, Z. Su and Y. Yu. (2008) Exploring folksonomy for personalized search. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08, pages 155–162, New York, NY, USA, ACM.

[Yanbe et al., 2007] Y. Yanbe, A. Jatowt, S. Nakamura and K. Tanaka. (2007): Towards improving web search by utilizing social bookmarks. In Proceedings of the 7th international conference on Web engineering, ICWE'07, pages 343–357, Berlin, Heidelberg, Springer-Verlag.

[Zaman and Li, 2014] N. Zaman, J. Li (2014) 'Semantics-enhanced Recommendation System for Social Healthcare'. In Proceedings of the IEEE 28th International Conference on Advanced Information Networking and Applications.

[Zaini et al., 2012] N. Zaini, M.F. Abdul Latip, H. Omar, L. Mazalan, H. Norhazman (2012) 'Online Personalized Audio Therapy Recommender based on Community Ratings'. In Proceedings of the International Symposium on Computer Applications and Industrial Electronics (ISCAIE 2012), December 3-4, 2012, Kota Kinabalu Malaysia.

[Zanardi and Capra, 2011] V. Zanardi, L. Capra. (2011): A Scalable Tag-based Recommender System for New Users of the Social Web. In: Proc. of the 2nd International Conference on Database and Expert Systems Applications.

[Zhang and Yang, 2014] M. Zhang, C.C. Yang (2014)' Classifying User Intention and Social Support Types in Online Healthcare Discussions'. In Proceedings of the EEE International Conference on Healthcare Informatics.

[Zhang et al., 2009] R. Zhang, Y. Chang, Z. Zheng, D. Metzler, and J.-y. Nie.(2009) Search result re-ranking by feedback control adjustment for time-sensitive query. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, NAACL-Short '09, pages 165–168.

[Zhao et al., 2008] S. Zhao, N. Du, A. Nauerz, X.Zhang, Q.Yuan, R. Fu, (2008): Improved recommendation based on collaborative tagging behaviors. In Proc. of the International Conference on Intelligent User Interfaces (IUI'08), pp. 413–416. ACM Press, Gran Canaria.

[Zhuang et al., 2006] L. Zhuang, F. Jing, X. Zhu, L. Zhang (2006) Movie review mining and summarization. In: Proceedings of the 15th ACM international conference on information and knowledge management. ACM, New York, pp 43–50.