THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Cotutelle internationale avec "Institut supérieur de Gestion de Tunis" Université de Tunis

---

Présentée et soutenue par :
Ameni KACEM SAHRAOUI

le mardi 13 juin 2017

Titre :

Personalized information retrieval based on time-sensitive user profile

---

École doctorale et discipline ou spécialité :

ED MITT : Image, Information, Hypermedia

Unité de recherche :

I.R.I.T

Directeur/trice(s) de Thèse :

Mohand BOUGHANEM et Rim FAIZ

Jury :

Lynda TAMINE Professeur, Université de Toulouse3 Présidente
Patrice BELLOT Professeur, Université d'Aix-Marseille Rapporteur
Chiraz LATIRI Professeur, Université de La Manouba Rapporteur
Sylvie CALABRETTO Professeur, Université de Lyon1 Examinateur
Mohand BOUGHANEM Professeur, Université de Toulouse3 Directeur
Rim FAIZ Professeur, Université de Carthage Directrice

This thesis is dedicated to
everyone who supported me!

Ameni

Abstract

Recently, search engines have become the main source of information for many users and have been widely used in different fields. However, Information Retrieval Systems (IRS) face new challenges due to the growth and diversity of available data. An IRS analyses the query submitted by the user and explores collections of data with unstructured or semi-structured nature (e.g. text, image, video, Web page etc.) in order to deliver items that best match his/her intent and interests.

In order to achieve this goal, we have moved from considering the query-document matching to consider the user context. In fact, the user profile has been considered, in the literature, as the most important contextual element which can improve the accuracy of the search. It is integrated in the process of information retrieval in order to improve the user experience while searching for specific information.

As time factor has gained increasing importance in recent years, the temporal dynamics are introduced to study the user profile evolution that consists mainly in capturing the changes of the user behavior, interests and preferences, and updating the profile accordingly. Prior work used to discern short-term and long-term profiles. The first profile type is limited to interests related to the user's current activities while the second one represents user's persisting interests extracted from his prior activities excluding the current ones. However, for users who are not very active, the short-term profile can eliminate relevant results which are more related to their personal interests. This is because their activities are few and separated over time. For users who are very active, the aggregation of recent activities without ignoring the old interests would be very interesting because this kind of profile is usually changing over time.

Unlike those approaches, we propose, in this thesis, a generic time-sensitive user profile that is implicitly constructed as a vector of weighted terms in order to find a trade-off by unifying both current and recurrent interests.

User profile information can be extracted from multiple sources. Among the most promising ones, we propose to use, on the one hand, searching history. Data from searching history can be extracted implicitly without any effort from the user and includes issued queries, their corresponding results, reformulated queries and click-through data that has relevance feedback potential. On the other hand, the popularity of Social Media makes it as an invaluable source of data used by users to express, share and mark as favorite the content that interests them.

First, we modeled a user profile not only according to the content of his activities but also to their freshness under the assumption that terms used recently in the user's activities contain new interests, preferences and thoughts and should be considered more than old interests. In fact, many prior works have proved that the user interest is decreasing as time goes by. In order to evaluate the time-sensitive user profile, we used a set of data collected from Twitter, i.e a social networking and microblogging service. Then, we apply our re-ranking process to a Web search system in order to adapt the user's online interests to the original retrieved results.

Second, we studied the temporal dynamics within session search where recent submitted queries contain additional information explaining better the user intent and prove that the user hasn't found the information sought from previous submitted ones. We integrated current and recurrent interactions within a unique session model giving more importance to terms appeared in recently submitted queries and clicked results. We conducted experiments using the 2013 TREC Session track and the ClueWeb12 collection that showed the effectiveness of our approach compared to state-of-the-art ones.

Overall, in those different contributions and experiments, we prove that our time-sensitive user profile insures better performance of personalization and helps to analyze user behavior in both session search and social media contexts.

Keywords: Personalized Search, User Profile, Freshness, Temporal Analysis, Social Media, Session Search

Résumé

Les moteurs de recherche, largement utilisés dans différents domaines, sont devenus la principale source d'information pour de nombreux utilisateurs. Cependant, les Systèmes de Recherche d'Information (SRI) font face à de nouveaux défis liés à la croissance et à la diversité des données disponibles. Un SRI analyse la requête soumise par l'utilisateur et explore des collections de données de nature non structurée ou semi-structurée (par exemple: texte, image, vidéo, page Web, etc.) afin de fournir des résultats qui correspondent le mieux à son intention et ses intérêts.

Afin d'atteindre cet objectif, au lieu de prendre en considération l'appariement requête-document uniquement, les SRI s'intéressent aussi au contexte de l'utilisateur. En effet, le profil utilisateur a été considéré dans la littérature comme l'élément contextuel le plus important permettant d'améliorer la pertinence de la recherche. Il est intégré dans le processus de recherche d'information afin d'améliorer l'expérience utilisateur en recherchant des informations spécifiques.

Comme le facteur temps a gagné beaucoup d'importance ces dernières années, la dynamique temporelle est introduite pour étudier l'évolution du profil utilisateur qui consiste principalement à saisir les changements du comportement, des intérêts et des préférences de l'utilisateur en fonction du temps et à actualiser le profil en conséquence. Les travaux antérieurs ont distingué deux types de profils utilisateurs : les profils à court-terme et ceux à long-terme. Le premier type de profil est limité aux intérêts liés aux activités actuelles de l'utilisateur tandis que le second représente les intérêts persistants de l'utilisateur extraits de ses activités antérieures tout en excluant les intérêts récents. Toutefois, pour les utilisateurs qui ne sont pas très actifs dont les activités sont peu nombreuses et séparées dans le temps, le profil à court-terme peut éliminer des résultats pertinents qui sont davantage liés à leurs intérêts personnels. Pour les utilisateurs qui sont très actifs, l'agrégation des activités récentes sans ignorer les intérêts anciens serait très intéressante parce que ce type de profil est généralement en évolution au fil du temps.

Contrairement à ces approches, nous proposons, dans cette thèse, un profil utilisateur générique et sensible au temps qui est implicitement construit comme un vecteur de termes pondérés afin de trouver un compromis en unifiant les intérêts récents et anciens.

Les informations du profil utilisateur peuvent être extraites à partir de sources multiples. Parmi les méthodes les plus prometteuses, nous proposons d'utiliser, d'une part, l'historique de recherche, et d'autre part les médias sociaux. En effet, les données de l'historique de recherche peuvent être extraites implicitement sans aucun effort de l'utilisateur et comprennent les requêtes émises, les résultats correspondants, les requêtes reformulées et les données de clics qui ont un potentiel de retour de pertinence/rétroaction. Par ailleurs, la popularité des médias sociaux permet d'en faire une source inestimable de données utilisées par les utilisateurs pour exprimer, partager et marquer comme favori le contenu qui les intéresse.

En premier lieu, nous avons modélisé le profil utilisateur utilisateur non seulement en fonction du contenu de ses activités mais aussi de leur fraîcheur en supposant que les termes utilisés récemment dans les activités de l'utilisateur contiennent de nouveaux intérêts, préférences et pensées et doivent être pris en considération plus que les anciens intérêts surtout que de nombreux travaux antérieurs ont prouvé que l'intérêt de l'utilisateur diminue avec le temps. Nous avons modélisé le profil utilisateur sensible au temps en fonction d'un ensemble de données collectées de Twitter (un réseau social et un service de micro-blogging) et nous l'avons intégré dans le processus de reclassement afin de personnaliser les résultats standards en fonction des intérêts de l'utilisateur.

En second lieu, nous avons étudié la dynamique temporelle dans le cadre de la session de recherche où les requêtes récentes soumises par l'utilisateur contiennent des informations supplémentaires permettant de mieux expliquer l'intention de l'utilisateur et prouvant qu'il n'a pas trouvé les informations recherchées à partir des requêtes précédentes. Ainsi, nous avons considéré les interactions récentes et récurrentes au sein d'une session de recherche en donnant plus d'importance aux termes apparus dans les requêtes récentes et leurs résultats cliqués. Nos expérimentations sont basés sur la tâche Session TREC 2013 et la collection ClueWeb12 qui ont montré l'efficacité de notre approche par rapport à celles de l'état de l'art.

Au terme de ces différentes expérimentations, nous prouvons que notre modèle générique de profil utilisateur sensible au temps assure une meilleure performance de personnalisation et aide à analyser le comportement des utilisateurs dans les contextes de session de recherche et de médias sociaux.

Mots-Clés: Recherche personnalisée, Profil Utilisateur, Fraîcheur, Analyse Temporelle, Médias Sociaux, Session de Recherche

Acknowledgements

I wish to express my immense gratitude to my supervisors and jury members for their encouraging and constructive comments and suggestions.

I want to thank my advisor *Pr. Rim FAIZ* for all for her encouragement and guidance that help the progression as well as her recommendations allowing me to improve this research. I would like to express my special gratitude to *Pr. Mohand BOUGHANEM* for accepting me among his team, introducing me the field of information retrieval and inspiring me throughout this research. I also want to thank him for his uninterrupted encouragement and efforts, which are things I will forever cherish.

My special thanks go to *Orange Tunisia Corporation*, *PASRI* and *ANPR* members that considered me worthy of the scholarship MOBIDOC. It surely would have been hard to complete the research conducted in this thesis without their support. In particular, I want to thank *Mr. Mohamed Arbi BEN YOUNES* and *Mrs. Asma ENNAIFER* as well as all the dream team DRE: *Youssef*, *Abdelaziz*, *Dhekra*, *Afef*, *Leila*, *Nizar*, *Aida*, *Belhassen*, *Asma*, *Bassem*, *Walid*, *Ayoub*, *Mariem*, *Emna*, *Salma*, *Amira*, *Lotfi* and without forgetting *Mehdi* and *Zbeida* and of course all *ODC* members.

I want also to thank all *ISG* and *LARODEC* professors and colleagues, especially *Rami*, *Dhouha*, *Sondess*, *Maha*, *Haithem*... and *IRIT* members especially: *Mohamed*, *Lamjed*, *Fatma*, *Baptiste*, *Thomas*, *Hung*, *Paul*, *Thibaut*, *Manel*, *Hela* and in particular *Bilel* and *Rafik* for their support and help during my internships in IRIT, as well as my dear and precious *Ghada* for everything she did for me.

My thanks go too to my dear friends *Nouha*, *Sarra*, *Maroua* as well as my companions on the road *Mariem* and *Imen* for all moments shared together and mostly for their support, love and for believing in me.

Lastly, and most importantly, I want to thank my affectionate and supportive family for its constant support:
I am highly grateful to my father *Habib* (I never thought that I would lose you so quickly- RIP) and to my mother *Leila*; their regular and unlimited efforts and encouraging words helped me overcome all encountered difficulties and achieve my goals in life as well as being who I am today. I could write pages expressing my gratitude and love. Thank you brother *Mohamed Ali* and sisters *Meryem* and *Feten* for always supporting me, helping me through anything in life and motivating me to go forwards to further success.
I am particularly grateful to my husband Ahmed who never complained and encouraged me to go abroad to study and work even if this meant being far away from him. He constantly cheered me when I was down and helped me when I needed advice. I am very blessed to have him near me.
My thanks to my husband's family for their continuous encouragement especially *Abdellatif*, *Fathia*, *Ichraf* (RIP my dear), *Yosra*, *Amine* and *Taha*.

I thank my exceptional family *Kacem* and also *Smida* family for always encouraging me to do better and improve.

I wish I can make you all proud through this Ph.D.!

List of Publications

- ## International Conferences

1. **A. Kacem**, M. Boughanem, and R. Faiz, *"Emphasizing Temporal-based User Profile Modeling in the Context of Session Search"*.
   In: The 32nd ACM Symposium on Applied Computing, SAC 2017, April 3-6, 2017, Marrakesh, Morocco. ACM, pp. 925–930.
   Paper Link

2. **A. Kacem**, R. Belkaroui, D. Jemal, H. Ghorbel, R. Faiz, I. Hammami Abid, *"Towards Improving e-Government Services Using Social Media-Based Citizen's Profile Investigation"*.
   In: Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance, ICEGOV 2016, Montevideo, Uruguay, March 1-3, 2016, pp. 187–190.
   Paper Link

3. **A. Kacem**, M. Boughanem, and R. Faiz, *"Time-Sensitive User Profile for Optimizing Search Personlization"*.
   In: User Modeling, Adaptation, and Personalization: 22nd International Conference on User Modeling, Adaptation and Personalization, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings. Ed. by Vania Dimitrova et al. Springer International Publishing, pp. 111–121.
   Paper Link

- ## National Conferences

1. D. Jemal, R. Belkaroui, **A. Kacem**, H. Ghorbel and R. Faiz, *"Towards Tunisia Smart City: Citizen Profile Investigation for Collaborative e-Government"*. In: 3rd Conference on E-Government Smart Tunisia and Smart Cities, E-GOV Tunisia, May 10th-13th, 2016, Tunis, Tunisia. Poster.

Contents

List of Figures

List of Tables

Introduction

Chapter 1

Introduction

1.1 Context and Problem Description

Information retrieval (IR) is a research discipline that integrates models and techniques which aim is to facilitate the access to relevant items for a user. It is defined as *"the techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system"* (*Merriam-Webster Dictionary*).

Recently, search engines have become the main source of information for many users and have been widely used for different fields. For example, Google, which is the top of the 15 most popular search engines with 1,600,000,000 unique monthly visitors as of November 2016 [1], has 7,766 billion average searches per day during the year of 2015 [2]. The observations of the Table 1.1[2] demonstrate that the annual average number of Google searches has increased by more than 100% from 2010 to 2015.

Year	Annual Number of Google Searches	Average Searches Per Day
2015	2,834,650,000,000	7,766,000,000
2014	2,095,100,000,000	5,740,000,000
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000
2010	1,324,670,000,000	3,627,000,000

TABLE 1.1: Google Search Statistics

However, Information Retrieval Systems (IRS) face new challenges due to the growth and diversity of available data (Allan et al., 2012). In fact, the

[1] http://www.ebizmba.com/articles/search-engines
[2] http://www.statisticbrain.com/google-searches/
[2] http://trec.nist.gov/pubs/trec22/appendices/session.html

amount of information available online is growing exponentially to become 4,76 billion pages as of November 2016[3]. Simultaneously, the number of Internet users has increased reaching over 3,5 billion as of September 2016.
Consequently, IR systems' users are often faced with too many irrelevant results because those systems do not take into consideration the context in which the query was submitted.
Moreover, although the same query is submitted by different users, they may have different intentions and goals (Dou et al., 2007) especially when the queries are short (Jansen et al., 2000) and ambiguous (Cronen-Townsend and Croft, 2002). Among famous examples of such ambiguous queries that are used in the literature, there are bass (fish or instrument), java (coffee, language programming or island), jaguar (Apple software, animal or car) and IR application (Infrared application or Information Retrieval application).

The personalization of information retrieval was proposed to overcome those challenges. The goal of personalized search engines is to provide search results that fit the individual user's information needs and match his/her interests instead of always providing the same results to a query regardless of the user who submitted it. In order to achieve this goal, we have moved from considering the query-document matching to consider the user context. In fact, the user profile has been considered as the most important contextual element which can improve the accuracy of the search (Park, 1994).

The user profile representation requires collecting information about the user that can be extracted from multiple sources. Among the used methods, the searching history attracts attention. Data from searching history can be extracted implicitly without any effort from the user and includes issued queries, their corresponding results, reformulated queries and click-through data that have relevance feedback potential.

Furthermore, we moved recently from using the Web as a mean of information access to generating data, a so called user generated content (UGC), especially with the increase of social services that emerged with the Web 2.0. In fact, those services, like tagging, microblogging, sharing media..., require users to be active and allow them expressing their interests, opinions and preferences, sharing and marking as favorite the interesting content. Thus, social media platforms are considered an invaluable source of data providing a better understanding of users' behavior and preferences (Cai and Li, 2010).

As time represents a challenging dimension nowadays in the field of information retrieval, considering the evolution of the user's interests represents a non-trivial task used to improve search performance.
In fact, it is beneficial to include a temporal factor in the user profile because content shared lately concerns the user more than content shared long time ago (Abel et al., 2011). This idea is underscored by the growth of Internet users and content, leading users to be more interested in recent content and data streams. Prior works regarding the evolution of user profile have introduced two temporal entities: short-term and long-term profiles. The

[3]http://www.worldwidewebsize.com/

short-term profile is limited to current activities whereas the long-term one includes the old interests ignoring the recent ones.
However, we find that the fact of discerning the short-term and long-term user profiles does not necessarily reflect the user's needs. For users who are not very active, the short-term profile can eliminate relevant results which are more related to their personal interests. This is because their activities are few and discontinuous over time. For users who are very active, the considering both the recent and old interests would be very interesting because this kind of profile is usually changing over time.

1.2 Research Questions and Goals

Search engines need to have deep knowledge about the user's activities and not be limited to the query's keywords in order to improve the search experience and to understand the user's intent. More precisely, we address the following research questions:

- **Which sources are considered sources of evidence for user profile modeling?**
  The explicit user profile is considered most of the time a burden. Users are not always willing to specify their personal information and fill in multiple forms. Creating the user profile implicitly is a solution allowing gathering more information about the user.
  Our goal is to extract interests of users without their involvement (Sugiyama et al., 2004; Tchuente et al., 2013). We aim at studying two sources: the searching history and the social media. As mentioned previously, those sources allow inferring users' interests thanks to collecting and analyzing their searching behavior and interactions as well as their published posts.

- **How do temporal feature affect the quality of user models in the context of personalized search ? How can current and recurrent interests be used in order to enhance personalization?**
  Many of state-of-the-art approaches assign more importance to the frequent terms no matter their moment of use. Time is often used to discern short-term and long-term user profiles. Short-term profile considers current actions while the long-term profile is built according to several previous actions.
  The fact of discerning the short- and long-term interests requires the use of a time interval that may include several interests (Dumais et al., 2003), or session's boundaries mechanisms where a session is defined by a set of queries related to the same information need (Shen and Zhai, 2003; Daoud et al., 2009).

1.3 Contributions

In this dissertation, the user profile is not only built according to the content of his activities but also to their freshness. We study the proposed profile within two contexts: the Web and Social Media.
The main contributions of this thesis consist on:

- *Modeling a time-sensitive user profile*
  We propose a generic user profile model based on time-sensitivity. We integrate the temporal dynamics into the user profile under the assumption that terms used recently in the user's activities contain additional information explaining better his intent, his interests, and his preferences... They should not be considered as old interests, as many prior works have proved that user's interests decrease as time goes by.
  This profile model can be used for both personalization and recommendation as it reflects the recurrent and current interests at different scales emphasizing newly expressed ones.

- *Studying social media-based user profiling*
  In this first context, we use Twitter as a source of data collection (Kacem et al., 2014; Kacem et al., 2016). Evaluating the user profile temporal dynamics within microblogging system is an interesting task because it allows users to (1) keep tuned with their friends and followers, (2) be updated with friends, family and followers, and (3) get instant feedback for publications (RkjIslam, 2016). Thus, the temporal dynamics are always related to microblogging due to its evolving and changing nature.

- *Studying the temporal dynamics in the context of searching history*
  We model a session profile based on user's submitted queries, results and clicks within a session. We adjust the users' used content using time-sensitivity in order to enhance recent interactions without ignoring the previous ones (Kacem et al., 2017).

For both contexts addressed, we evaluate the performance of our proposed models using two testbeds: the first one is a collection extracted from Twitter and the second one is the ClueWeb12 collection[4]. The experiments conducted using those data sets have shown the effectiveness of our user profile modeling approach based on freshness compared to state-of-the-art approaches.

1.4 Thesis structure

This dissertation is organized into two parts. The first one presents the context in which our work is carried out, which is personalized search using temporal user profile. The second part details our contribution.
The aim of the first part entitled ***From Standard to Temporal Information Retrieval*** is to give an overview of the process of information retrieval,

[4]http://lemurproject.org/clueweb12/

state-of-the-art approaches of the personalization and the emergence of temporal information retrieval.

- In *Chapter* 2, we begin with studying the basic concepts, the process and the models of IR and we present the evaluation of IR models. Then, we discuss the challenges that modern IR systems are facing, in particular those related to Web data and the user's needs. Finally, we cite the most used measures and campaigns to evaluate IR models.
- In *Chapter* 3, we review related works focusing on existing personalized search types, techniques and approaches, and on the user profiling approaches currently in use. Particularly, we give attention to user profile modeling in the scope of social media.
- In *Chapter* 4, we study the emergence of temporal information retrieval, and discuss its integration in the user profiling in two aspects: *(1)* the Web especially Session Search *(2)* Social Media.

Part two entitled ***On Using Time-Sensitive User Profiling for Personalized Search*** is dedicated to describe our thesis contributions and the main experiments proposed to evaluate our profiling strategy.

- In *Chapter* 5, we propose a time-sensitive profile extracted from social media and more precisely from Twitter. We compare our model with state-of-the-art ones and prove the effectiveness of integrating a temporal feature in addition to the frequency.
- In *Chapter* 6, we present an effective approach to personalize a current query in a session. This approach is based on understanding the user's interests and preferences modeled through a user profile and expressed in his previous interactions such as submitted queries, reformulated queries and clicked results. We used TREC Session 2013 as a framework to conduct experiments providing test collections and evaluation measures in order to compare our temporal model over baselines.
- In *Chapter* 7, we present the relevance of the time-sensitive user profile in Web and mobile contexts within Orange Tunisia Corporation as our research work is carried out within the MOBIDOC device, under the PASRI program[5], administered by the ANPR[6], and funded by the European Union[7] and Orange Tunisia Corporation[8].

Chapter 8 concludes this thesis, summarizes our findings, and opens new perspectives.

[5] http://www.pasri.tn/
[6] http://www.anpr.tn/
[7] http://europa.eu/
[8] https://plus.orange.tn

Part I

From Standard to Temporal Information Retrieval

Chapter 2

Background of Information Retrieval

2.1 Introduction

Information retrieval (IR) is a computer science discipline that integrates models and techniques which aim is to facilitate the access to relevant items for a user.

Search engines represent the most known IR application. They have become the main source of information for many users and have been widely used in different fields. However, Information Retrieval Systems (IRS) face new challenges due to the growth and diversity of data and users. In fact, an IRS analyses the query submitted by the user and explores collections of data with unstructured or semi-structured nature (e.g. text, image, video, Web page etc.) in order to deliver items that best match his/her intent and interests.

In order to achieve this goal, IRS have moved from the query-document matching to consider also the user context in the retrieval process. Before

presenting approaches considering the context of the user and the personalization of search results, we introduce in this chapter a general overview of the IR field.

First, we define Information Retrieval and present the basic concepts of the IR field: document, query and relevance. Section 2.3 describes the process of IR with its steps: indexing and matching, as well as the relevance feedback. Section 2.4 outlines the main IR models while Section 2.5 describes how to evaluate them. Afterward, we provide an overview of how to evaluate IR systems and approaches.

2.2 Basic Concepts

We quote below the definitions of IR given in its original forms:

- *"an information system, that is, a system used to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations."* (Salton and McGill, 1986).
- *"The user expresses his information need in the form of a request for information. Information retrieval is concerned with retrieving those documents that are likely to be relevant to his information need as expressed by his request"* (Van Rijsbergen, 1986).
- *"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)"* (Manning et al., 2008).
- *"the techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system"* (*Merriam-Webster Dictionary*).

From these definitions, we conclude that an Information Retrieval System (IRS) is a software tool for data representation, storage and search. The selected documents are relevant when they meet the user's need expressed in the submitted query. We, then, define the following key concepts:

- **Document**
  A Document is known as any unit of information that can be an answer to a need expressed by the user. A document can be a text, a piece of text, image, video, URL, etc.
- **Query**
  A query is a representation of the user's information need. For a given search session, the user must submit a query to the search engine that describes through keywords what information he is looking for.
- **Relevance**
  According to preliminary definitions (Saracevic, 1970), relevance measures how a document matches a query. The concept of relevance is the primary criteria for evaluating IRS.

In addition, many studies have been conducted around relevance which is often shown as the degree of similarity between the query's representation and the document content (Boughanem and Savoy, 2008).

Saracevic presented a critical review of the nature of relevance (Saracevic, 1996) and defined five types of it :

- *The algorithmic relevance (or system)*: It is often identified by a score of similarity between document's content and the query's terms.
- *The topical relevance*: It reflects the quality of matching between the information of retrieved documents and the topic expressed in the query's keywords.
- *The cognitive relevance*: It is defined as the relationship between the knowledge state of the user and the information carried by the documents.
- *The situational relevance (utility)*: It is the relevance of the resources according to the situation, task, or problem of the user.

2.3 Standard IR Process

The main purpose of IRS is to select all relevant documents which best match the user's information needs. For decades, information retrieval was used by professional searchers, but nowadays hundreds of millions of people are using information retrieval tasks daily.

Bringing out such a system consists mainly in implementing a process, shown in Fig. 2.1 which lies in two main phases: indexing and matching.

2.3.1 Indexing

Indexing is a set of techniques employed to transform the documents (or queries) into substitutes or descriptors which can represent their contents (Salton and McGill, 1986). These descriptors characterize the indexing language which is often represented through a structure. This structure is based on a set of keywords or phrases representing the textual content of the document.

Therefore, the indexing consists in detecting the most representative terms of the document's content. Indexes can be carried out on three forms: manual, automatic or semi-automatic indexing.

- **Manual Indexing**: An expert in the field chooses the terms that he considers relevant in describing the semantic content of the document by using a controlled vocabulary (hierarchical list, thesaurus[1], dictionary...).

[1] A thesaurus is an organized list of descriptors (keywords) grouped together according to semantic relations (synonyms or antonyms).

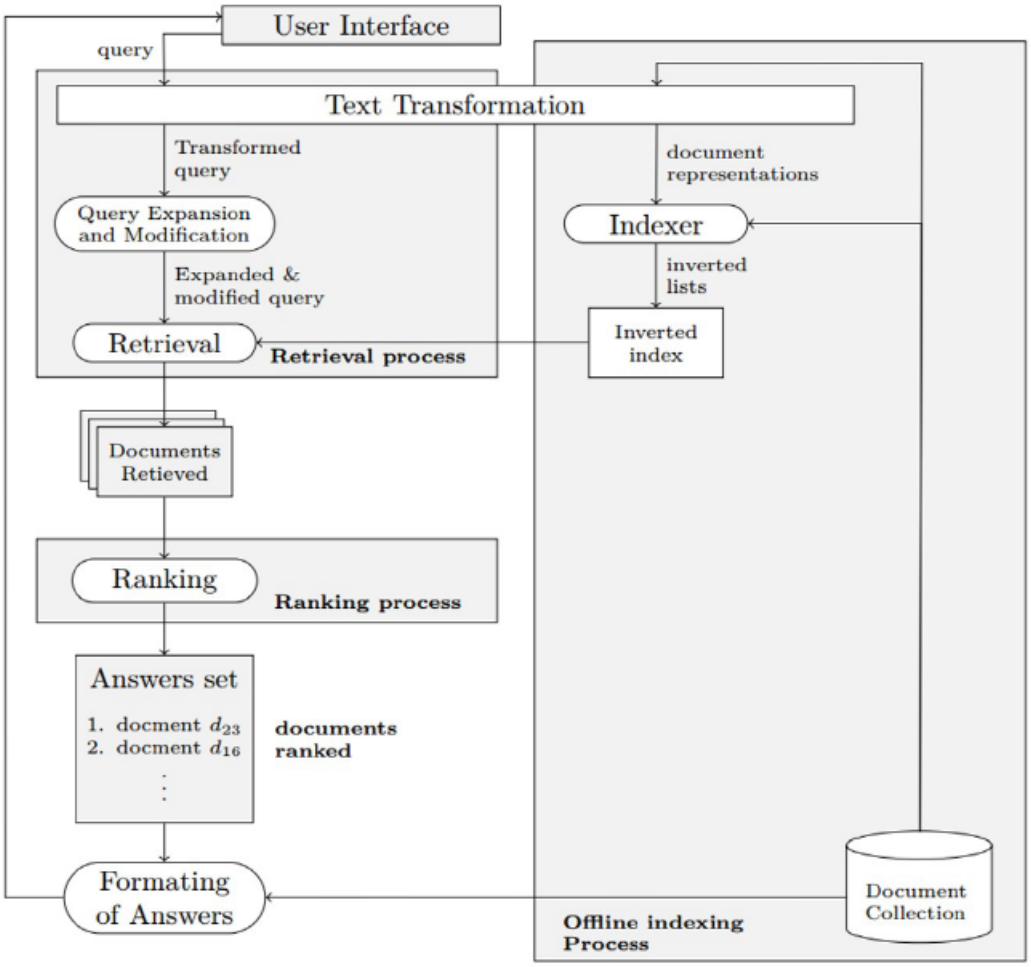

FIGURE 2.1: General Model of the IR Process (Baeza-Yates and Ribeiro-Neto, 1999)

- **Automatic Indexing**: This type of indexing does not involve experts. The indexing process is completely automated based on algorithms associating descriptors to documents.
- **Semi-automatic Indexing**: It is a combination of the two previous methods: the automatic process allows the extraction of the document's terms. However, the final choice of descriptors is left to the expert, which uses a controlled vocabulary or thesaurus.

The task of indexing involves not only deriving which parts and keywords are extracted from the documents in the collection, but also a complex computational process: Tokenization, Stopwords removal, Stemming and Weighting of words.

Tokenization

During this phase, all remaining text is parsed. A set of words is extracted using a simple lexical analysis to identify words by recognizing spaces separating words, special characters, numbers, punctuation, etc..

Stopwords removal

The stopwords removal refers to eliminating very frequent terms, which are mostly prepositions and articles as "in", "a", "the", "to" or pronouns like "I", "he", "she", "it" for example. In fact, these words are likely to be not relevant to describe a document's content and therefore can be filtered out. Ruling out these words from the index can result in significant space savings, while not affecting retrieval performance.

Traditionally, stopwords are chosen from a fixed list, which is constructed from lexical information. This has led to the appearance of different stopwords lists which vary in size, like Smart's list, (Buckley et al., 1995), comprised 571 terms, or Okapi's one (Robertson and Walker, 2000), made up of 220 terms.

Stemming

Stemming refers to the conflation of words to their lemmatized base form, where the morphological variants of words are stripped to one single suffix entry. Thus,"computer", "computing", "compute" is reduced to "compute". This process employs shallow linguistic information (e.g. removing endings of nouns), as well as language pattern matching rules (Porter, 1997).

Weighting

The weighting is one of the important steps in IR. This step is often related to IR models and used to assign weights to terms. Weighting can be local *(TF)*, global *(IDF)* or both *(TF.IDF)*.

- **Term Frequency (TF):** If local weights are used, then term weights are generally expressed as term frequencies TF (Manning et al., 2008). The underlying idea is that the more frequent a term is, the more important it is in the document's representation. Let's assume that d_j is a document and t_i is a term, then the frequency TF_{ij} of the term in the document is often used directly or expressed as one of the following froms:

$$TF_{ij} = 1 + log(td_{ij}), TF_{ij} = \frac{td_{ij}}{\sum_k td_{kj}} \tag{2.1}$$

  where td_{ij} is the number of occurrences of term t_i in the document d_j. The denominator is the number of occurrences of all terms in the document d_j. The last variation normalizes the term frequency to avoid bias related to the length of the document.

- **Inverse Document Frequency (IDF):** Using the global weights, the weight of a term is specified by *IDF* (Manning et al., 2008). The main idea is to measure the overall representativeness of the term with respect to the collection of documents. However, rare terms have high IDF, contrary to frequent terms.

$$IDF_i = log\frac{N}{n_i}, IDF_i = log\frac{N - n_i}{n_i} \tag{2.2}$$

where N is the size of the collection (number of documents) and n_i is the number of documents containing the term i.

- **Term Frequency-Inverse Document Frequency (TF.IDF):** In this type of weighting, both the local and global weights are used. This is commonly referred to as TF.IDF weighting (Robertson and Sparck Jones, 1988):

$$TF.IDF = log(1 + TF) * IDF \tag{2.3}$$

One of the most currently used formulas in the field of IR is the formula of Okapi BM25 (Robertson et al., 1995) as the weight of term i in document j (denoted $w_(i, j)$) is given by:

$$w_{i,j} = log(\frac{N - n_i + 0,5}{n_i + 0,5}) * \frac{tf_{i,j} * (k_1 + 1)}{tf_{i,j} + k_1 * ((1 - b) + b * \frac{dl}{avgdl})} \tag{2.4}$$

where N is the size of the collection, n_i the number of documents containing the term t_i, dl is the size of the document d_j, $avgdl$ is the average sizes of documents in the collection and k_1, b are parameters that depend on the collection and the types of queries. The constants are usually chosen as: k_1 between 1:0 and 2:0, b equaling 0:75.

2.3.2 Document-Query Matching

The IRS includes a search process for selecting the information deemed relevant to the user. It involves a process of user's interaction with the IRS which includes the following steps:

First, the user expresses his information need in the form of a query. Second, the system creates the query's index which will be compatible with the model of documents' index. Finally, the system evaluates the match between the query and the documents.

To achieve the previous step, the IRS calculates a relevance score between the query and the indexed descriptors of the document's collection. The documents are then ranked acroding to their relevance score denoted RSV (Q,D) (Retrieval Status Value) where Q is a query and D is a document in the collection. The matching process is closely linked to the process of indexing and term weighting. In fact, there are two matching models:

- **Exact-match retrieval model:** The result is a list of documents respecting exactly the submitted query according to specific criteria. However, the documents returned are not sorted (Salton, 1971).

- **Best-match retrieval model:** The result is a list of documents supposed to be relevant to the query. The returned documents are sorted according to the relevance score (Robertson and Sparck Jones, 1988).

2.3.3 Query Reformulation

The reformulation of the initial query submitted by the use allows to change the terms expressing the user's information need during a session search. This step can be performed either manually or automatically.

- **Manual Reformulation** : can be performed when the user change himself the query's terms in order to submit a new one.
- **Automatic Reformulation** : can be obtained when the IR system uses the top-K relevant documents in order to extract the best terms and reformulate the query accordingly.

2.4 Overview of IR models

The models are important in the IR process. In fact, they allow guiding search and providing a formal description of the IR process and a theoretical framework for modeling the relevance measure. In addition, models can serve as a blueprint to implement an actual IR system. According to (Baeza-Yates and Ribeiro-Neto, 1999), an IR model is defined by a quadruple $(D, Q, F, R(q_i, d_j)$, where:

- D is a set of documents,
- Q is a set of queries,
- F is a modeling framework for D, Q, and the relationships among them,
- R(q_i, d_j) is a relevance or similarity function which ranks the documents with respect to a query.

Many models have been proposed in the information retrieval literature, but the selection made in this part gives a comprehensive overview of the basic types of modeling approaches.

2.4.1 The Boolean Model

The Boolean model (Salton, 1968) is the first model of information retrieval which is based on set theory and Boolean algebra. In this model, each document is represented by a logical conjunction of unweighted terms which represent the index of the document.

Indeed, the Boolean model considers that the index terms are present or absent in a document. Hence, the weights of terms in the index are binary, i.e either 0 or 1. Using the operators of Boole's algebra, query terms and their corresponding sets of documents can be combined to form new sets of

documents. The result of this function is a binary score written as follows: $RSV(q_i, d_j) = 1, 0$.

2.4.2 The Vector Space Model

The basic idea of the Vector Space Model (VSM) is to represent documents and queries as vectors in a multidimensional space, whose dimensions are the terms used to build an index representing the documents (Salton and McGill, 1986). Both of the documents and the query are represented as vectors: $\overrightarrow{d_j} = (w_{1,j}, w_{2,j}, ..., w_{n,j})$ and $\overrightarrow{q_i} = (w_{1,i}, w_{2,i}, ..., w_{n,i})$, where $w_{k,j}$ denotes the weight of a term t_k in a document d_j and $w_{k,q}$ denotes its weight in the query q_i.

The relevance is interpreted as a similarity measure such as *Jaccard* (Equation 2.5), *Sørensen–Dice* (Equation 2.6) or the *Cosine function* (Equation 2.7).

$$RSV(\overrightarrow{q_i}, \overrightarrow{d_j}) = \frac{\left|\overrightarrow{q_i} \cap \overrightarrow{d_j}\right|}{\left|\overrightarrow{q_i} \cup \overrightarrow{d_j}\right|} = \frac{\left|\overrightarrow{q_i} \cap \overrightarrow{d_j}\right|}{|\overrightarrow{q_i}| + \left|\overrightarrow{d_j}\right| - \left|\overrightarrow{q_i} \cap \overrightarrow{d_j}\right|} \tag{2.5}$$

$$RSV(\overrightarrow{q_i}, \overrightarrow{d_j}) = \frac{2.\left|\overrightarrow{q_i} \cap \overrightarrow{d_j}\right|}{|\overrightarrow{q_i}| + \left|\overrightarrow{d_j}\right|} \tag{2.6}$$

where $|\overrightarrow{q_i}|$ and $\left|\overrightarrow{d_j}\right|$ denotes the number of terms in query q_i and the document d_j respectively, $\left|\overrightarrow{q_i} \cap \overrightarrow{d_j}\right|$ represents the number of common terms between the document and the query.

$$RSV(\overrightarrow{q_i}, \overrightarrow{d_j}) = Cos(\overrightarrow{q_i}, \overrightarrow{d_j}) = \frac{\overrightarrow{q_i}.\overrightarrow{d_j}}{\|\overrightarrow{q_i}\| \, . \left\|\overrightarrow{d_j}\right\|} \tag{2.7}$$

where $\|\overrightarrow{q}\|$ and $\left\|\overrightarrow{d_j}\right\|$ denote euclidean norms of a query $\overrightarrow{q}$ and a document $\overrightarrow{d_j}$ vectors .

In fact, the more similar the two vectors are, the smaller the angle is, and the larger the cosine of this angle is. In addition, the VSM assigns a high ranking score to a document that contains only a few of the query terms if those terms occur occasionally in the collection but frequently in the document. This model makes the following assumption: The more similar a document-vector is to a query-vector, the more significant the document is to the query.

Unlike the Boolean model, the matching function evaluates a best-match between a document and a query, allowing users to find documents that satisfy approximately the query. In addition, the results can be ranked in decreasing order of similarity. However, this model assumes that the terms used to define the dimensions of the space are orthogonal or independent which is not realistic because relations among terms may exist.

2.4.3 The Probabilistic Models

(Maron and Kuhns, 1960) introduced the first probabilistic approach to information retrieval. This model states that an information retrieval system is supposed to rank the documents based on their probability of relevance to the query. Besides, the principle assumes that there is uncertainty in both the information need and the documents representations. One of the most effective probabilistic model based on the probabilistic ranking principle (Equation 2.8), is the BM25 model proposed by (Robertson and Sparck Jones, 1976).
This model uses the statistical distribution of the terms in both the relevant and non-relevant documents. For a given query q_i, a document d_j is selected if the probability that d_j is relevant, denoted by $P(R \mid d_j)$, is greater than the probability that d_j is not relevant, denoted by $P(\overline{R} \mid d_j)$, where R is the relevance event and $\overline{R}$ is the irrelevance event. The similarity score *RSV* (q_i, d_j) is given by:

$$RSV(q_i, d_j) = \frac{P(R \mid d_j)}{P(\overline{R} \mid d_j)} \tag{2.8}$$

Such Model is a probability-based model that is used in different fields such as spell correction, speech recognition, and question-answering (Jurafsky, 2012). This model (Ponte and Croft, 1998) is based on a representation of a document by computing the distribution of its terms/sentences and producing the query given this representation. Thus, instead of measuring the probability $P(R \mid d_j)$ of a document d_j to be relevant to a query q_i, this approach proposes a document model M_{d_j} and ranks documents according to the probability $P(q_i \mid M_{d_j})$ (Manning et al., 2008)

The relevance is then computed as follows:

$$RSV(q_i, d_j) = P(Q \mid M_{d_j}) = \prod_{t \in q_i} P(t \mid M_{d_j}) \tag{2.9}$$

where for each term *t* in the query q_i, the probability $P(t \mid M_{d_j})$ is measured based on the Maximum Likelihood. It is obtained through the frequency of a query's term t in the document d_j. However, this can lead to a null value when the query's term doesn't exist int the document. To overcome this drawback, techniques of smoothing were proposed. Smoothing refers to *"the adjustment of the maximum likelihood estimator of a language model so that it will be more accurate"* (Zhai and Lafferty, 2004). Among smoothing approaches, we cite: Jelinek-Mercer (Jelinek and Mercer, 1980), Dirichlet (MacKay and Peto, 1994), Additive or called also Laplace (Manning et al., 2008)...

2.5 Evaluation of IR systems

One meaning of evaluation is to *"ascertain the value or amount of something or to appraise it"* (Kiewitt, 1979). Mainly, evaluation of IR systems consists

of checking how effectively a system can find and introduce relevant information to the user.

The aim of a search engine is to achieve user satisfaction which can be estimated by appropriate success criteria usually measured by, for instance, the user satisfaction, the quality of results, and the run time...

In this section, we describe the evaluation based on test collections and metrics used to IR systems effectiveness.

2.5.1 Test Collections

Test collection is a laboratory testbed reflecting the real world. It allows having a static set of documents and a set of known relevant documents thanks to trained persons called *assessors* as described in Fig. 2.2.

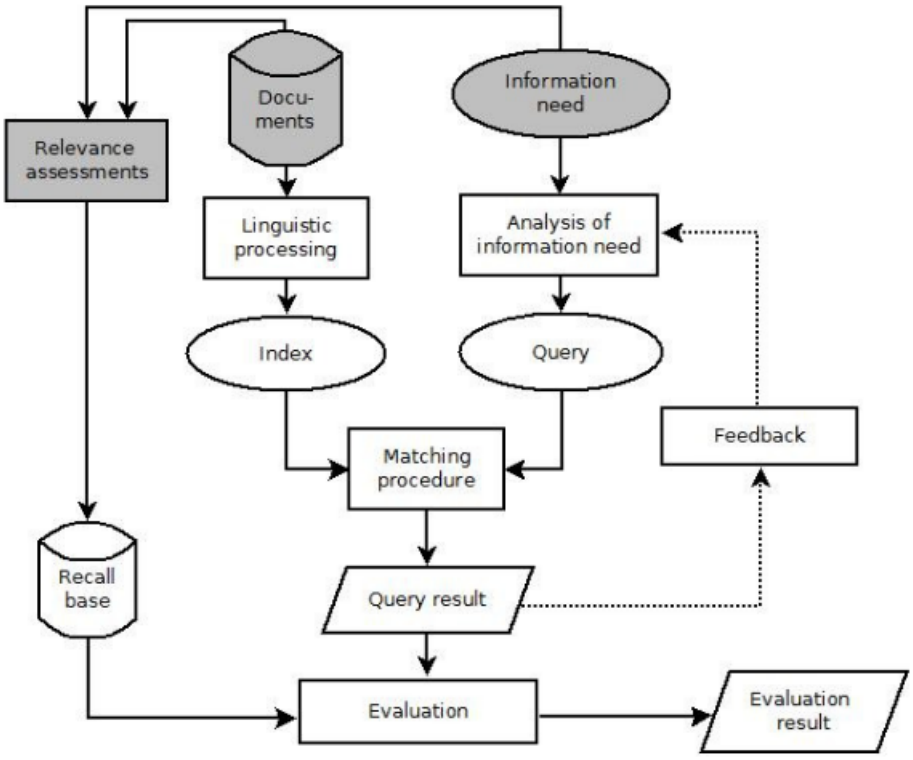

FIGURE 2.2: Evaluation Protocol using Test Collections (Heppin, 2012)

This methodology uses the Cranfield paradigm introduced by (Cleverdon et al., 1966) and started in the early 1960s. It resides in a common framework allowing to compare different IR systems, search strategies and search algorithms based on test collections composed of a set of queries, a set of documents and relevance judgements:

- *Queries*: The queries represent the description of the information needs and the expression of a *topic* that is processed by an information retrieval system. In order to get significant results, a set should be composed of at least 25 topics (Buckley and Voorhees, 2000).

- *Documents*: A set of documents is preselected based on their match to the topics.
- *Relevance Judgement*: Also called *Ground Truth*, relevance judgments identify documents that are relevant to a query. A gradual relevance score may be associated for each document-query pair. In an ideal test collection, all documents are assessed. However, Relevance Judgement is not a trivial task and it is often considered annoying as it requires time and effort from assessors especially for a large number of topics and documents.
  Thus, *pooling* (Jones et al., 1975) was used in TREC tasks. It consists of the following steps:
  - Top-K results from the rankings obtained by different search engines (or retrieval algorithms) are merged into a *pool*,
  - Duplicates are removed,
  - Documents are presented in a random order to the relevance judges (assessors).

The most famous collections are: TREC, NTCIR and CLEF.

- **TREC**[2]: The Text REtrieval Conference supports research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.
  Among TREC tracks we cite: *Session Search, Microblog, Web, Question Answering...*
- **NTCIR**[3]: NII Testbeds and Community for Information access Research is a series of evaluation workshops designed to promote research in Information Access technologies including information retrieval, question answering, text summarization, extraction, etc.
  It is focused on the specific language technologies necessary for Asian languages and cross-lingual searches among these languages and English.
- **CLEF**[4]: The Conference and Labs of the Evaluation Forum (formerly known as Cross-Language Evaluation Forum) is an initiative developed as a continuation of the Cross Language Track at TREC.
  Its main mission is to promote research, innovation, and development of information access systems based on multilingual information. Among CLEF tasks (called Labs) we find : *CLEF eHealth, Question Answering, Social Book Search, News Recommendation Evaluation Lab (NEWSREEL)...*

To perform the comparison of relevance provided by different approaches using these frameworks, relevance measures are calculated for all the topics of a specific track. In the following Section, we introduce in details the most commonly used measures.

[2]http://trec.nist.gov/
[3]http://research.nii.ac.jp/ntcir/index-en.html
[4]http://www.clef-initiative.eu/

2.5.2 Evaluation Metrics

A definition of an evaluation metric is given by (Pehcevski and Piwowarski, 2009): *"An evaluation metric is used to evaluate the effectiveness of information retrieval systems and to justify theoretical and/or pragmatical developments of these systems. It consists of a set of measures that follow a common underlying evaluation methodology"*.

Typically, the measures require a collection of documents and a query, whereas every document is either relevant or non-relevant to a particular query. In the following we explain some of the most common measures used in information retrieval.

Set-based Measures

According to Fig. 2.3, let R be the subset of relevant documents with respect to a query Q, and S be the selected search results set.
Using those features, we introduce, here, Precision, Recall and F-Measure metrics.

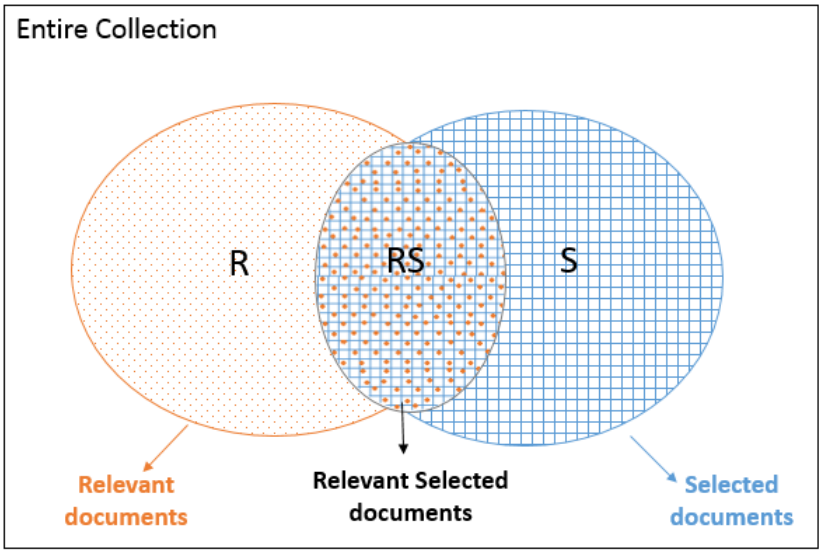

FIGURE 2.3: Relation between Relevant and Selected Documents

- **Precision** is the portion of the retrieved documents that are relevant to the user's information need. In fact, it shows the ability of a system to select all relevant documents in the collection.
  Then, it is given by the amount of relevant selected documents divided by the total number of selected documents:

$$Precision = \frac{|RS|}{|S|} \tag{2.10}$$

- **Recall** in information retrieval is the part of successfully retrieved documents that are relevant to the query. Indeed, it emphasizes the

ability of a system to select only relevant documents.
It is given by the ratio of relevant selected documents divided by the number of documents that are relevant to the query:

$$Recall = \frac{|RS|}{|R|} \tag{2.11}$$

- Usually there is a trade-off between precision and recall, i.e., the higher recall gets, the lower precision tends to be. Thus, a retrieval system is distinguished by the ratio of precision to recall called the F-Measure or F-Score:

$$F - Mesaure = \frac{2.Precision.Recall}{Precision + Recall} \tag{2.12}$$

By computing the precision and the recall at every position in the ranked sequence of the documents, one can plot a precision-recall curve, as illustrated in Fig. 2.4, plotting precision *P(r)* as a function of recall *r*.

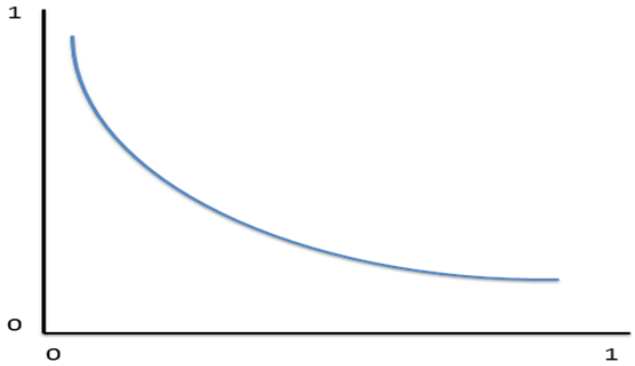

FIGURE 2.4: Aspect of a Precision-Recall Curve

Rank-based Measures

Several ranking measures have been proposed in the IR field in order to evaluate rank-based search systems. We present the popularly used ones: MAP, NDCG and MRR.

- **MAP@r** *Mean Average Precision* : For an evaluation query q_i, it measures the average over the precision values computed at each point in the ranking where a relevant document occurs.
  Then, using a set of evaluation queries Q we have the following expression:

$$MAP = \frac{1}{Q} \sum_{q_i \in Q} \frac{1}{r} \sum_{r \in R} Precision(q_i)@R \tag{2.13}$$

  Where Q is the number of queries, r represents the number of relevant documents for a query q_j and R is the rank of a relevant document.

- **NDCG@r** *Normalized Discounted Cumulative Gain*: Its main novelty is its ability to model different relevance levels (Järvelin and Kekäläinen,

2002). The DCG is a measure that gives more weight to highly ranked documents and allows the incorporation of different relevance levels. The *Discounted Cumulative Gain* is measured for each query q_j at the n^{th} position :

$$DCG_j^n = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{log_2 i} \tag{2.14}$$

Where rel_i is a relavance function assigned to the i^{th} document. The *NDCG* can be estimated from the *DCG* applied to the perfect ranking of relevance judgments according to their degree, denoted $IDCG_j^n$:

$$NDCG@n = \frac{\sum_{q_j \in Q} DCG_j^n}{\sum_{q_j \in Q} IDCG_j^n} \tag{2.15}$$

- **EER@s** *Expected Reciprocal Rank*: It is based on the cascade model of search that assumes a user scans through ranked search results in order to evaluate whether the document satisfies the query of each session, and if it does, stops the search (Chapelle et al., 2009):

$$ERR = E(1/s) = \sum_{k=1}^{K} \frac{1}{k} p(q, d_k) \prod_{i=1}^{k-1} (1 - p(q, d_i)) \tag{2.16}$$

where *s* denotes the rank at which we stop, *q* is a query in a session, *K* is the number of returned documents, where the probability that document *k* satisfies the user query is given by the transform of the editorial grade assigned to the query-document pair : *p(q, dk)*.

2.6 Conclusion

In this chapter, we introduced the field of Information Retrieval (IR) and its main concepts. We also presented main state-of-the-art models and evaluation approaches and systems.

The major limitation of almost of traditional search systems is returning the same list of results to a same query submitted by different users. However, users may have different interests and therefore have different information needs. Thus, they may have to go through many irrelevant results or try several queries before finding the desired information.

Personalization has become more developed with IR challenges of access to information such as finding relevant information in a diverse area with a considerable size and a variety of users.

In the next chapter, we carry out our specific analysis of the state-of-the-art works regarding personalized search and user profiling.

Chapter 3

User Profiling for Personalized Search

3.1 Introduction and Context

From the general overview of Information Retrieval (IR) described in Chapter 2, we note that there is a need to integrate the user dimension in the retrieval process in order to facilitate finding relevant data. Such integration is referred as *Personalized Information Retrieval*. The goal of personalized IR systems is to adapt results to the individual user, his/her preferences, needs, competence, background, knowledge and context.

The user profile represents the most important contextual element which can improve the accuracy of the search (Park, 1994). It is integrated in the process of IR in order to improve the user experience while searching for specific information.

User profile information can be extracted from multiples sources. Among the most promising ones, the browsing history attracts attention. Data from browsing history can be extracted implicitly without any effort from the user and includes issued queries, their corresponding results, reformulated

queries and click-through data that has relevance feedback potential (Snášel et al., 2010).

With the expansion of the Web 2.0 in recent years and the emergence of social media systems, the user can share content and be connected to other people such as family and friends. This innovation gave the user the chance to move from being a passive user who consumes content towards an active producer who creates content, shares it to be accessible to his community and/or modifies it to add his personal touch. Publications posted or edited by the user are called *User Generated Content (UGC)*. UGC are used by personalized IR systems in order to find relevant information about the user and its preferences and habits.

Among the most popular social media systems, we cite Twitter that has over 190 million of unique Twitter site visitors every month and 2.1 billion of Twitter search engine queries every day [1]. On Twitter, users express their interests, preferences, opinions, states... in a form of a 140-word-post called *"Tweet"*. Recent works take advantage of Twitter and other microblogging systems, socials networks (Facebook[2]), media Websites (Flicker [3])... as relevant data that provide a better understanding of users' behavior and preferences improving, therefore, personalized search results.

In this chapter, we discuss first the personalization of information retrieval and the integration of user profile through different approaches in the process of IR. In Section 3.3, we present the user modeling sources and representation forms. Then, we move to discussing, in Section 3.4, the emergence of the Social Web and its impact on information retrieval and user modeling.

3.2 Personalization of Information Retrieval

The standard information retrieval is based mainly on measuring the relevance of a document to a query according to fixed criteria or on the exploitation of link structure between documents.

However, most of IR approaches return the same list of search results based on the query but pay no attention to the users' specific interests and/or search context, a so called one-size-fits-all approach (Lawrence, 2000). Furthermore, due to some problems, such as individual differences in information needs, a user may have to go through many irrelevant results or submit several queries before finding the desired information.

As a result, information retrieval systems face a difficult challenge: providing search results that fit the individual user's information needs and match his/her interests instead of providing the same results to a query for all users. Consequently, we need to learn more about the user and its context in order to understand better those changes.

[1] http://www.statisticbrain.com
[2] www.facebook.com
[3] www.flicker.com

3.2.1 Personalized Search Systems

A Personalized IRS (PIRS) is a system that integrates the user into the information access unfolding. The PIRS is not only limited to model the user profile but it must be able to infer his intention when performing the search.

Data that allow representing the user profile can be extracted from multiple sources especially online social data as we discuss in Section 3.4 where users leave evidence of their interests and preferences.

Furthermore, PIRS must include:

- Techniques and tools for collecting personal user's information,
- Techniques and approaches to model the user profile, i.e representing his/her interests, preferences, knowledge and goals,
- A method to track the evolution of the profile and to update its components,
- Mechanisms and algorithms to integrate the user's profile in the process of information retrieval.

Results personalization can be achieved through three models: integration in the retrieval process, post-search personalization and pre-search personalization. Each of those models is described in detail in the following sections. All models are presented in Fig. 3.1 proposed by (Micarelli et al., 2007).

- **Integration in the retrieval process**
  The personalization is a unified process where user profiles are integrated to weight content.
- **Post-search personalization**
  This type of personalization (Micarelli and Sciarrone, 2004; Speretta and Gauch, 2005) consists of filtering and re-ranking relevant results that are initially returned from a standard search engine. The relevance feedback, in this type, is used to adjust weights of index terms in order to obtain relevant results for a specific user.
- **Pre-search personalization**
  This type of personalization is usually known as "query expansion" that we explain in detail in Section 3.3.4. In fact, popular terms, similar terms and related terms can be used, respectively, to introduce context, to resolve indexer-user mismatch and to clarify ambiguity.

These approaches requires learning the user as part of the search context in order to integrate his profile in the search process.

3.2.2 Context and User Profile

Many researches integrated the context and the user profile in the retrieval process. In the following sections we introduce both the concepts of context and user profile.

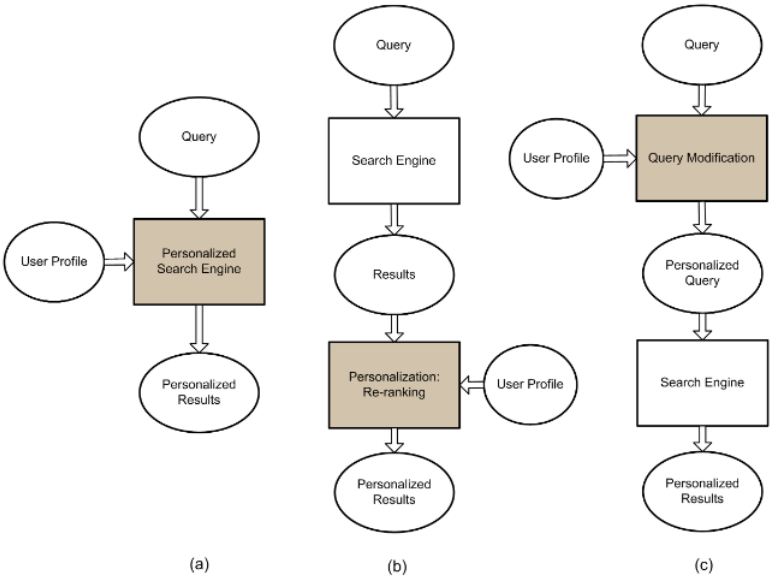

FIGURE 3.1: Personalization process where the user profile occurs during the retrieval process(a), in a distinct re-ranking activity (b) or in a pre-processing of the user query (c) (Micarelli et al., 2007)

- **Context**
  There have been several definitions of context proposed in the personalized IR literature that differ essentially in its components. For example, (Abowd et al., 1999) propose their own definition that covers many authors' definitions: *"Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves"*.

  Besides, contextual retrieval was defined by (Allan et al., 2003) as to *"Combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs"*. They considered three dimensions: *social context, work context and time*.

  In addition to the definitions of *"context"* and *"contextual retrieval"*, many authors tend to categorize the context by identifying classes and types (Ingwersen and Järvelin, 2005; Daoud et al., 2009). In fact, as the context is a multi-dimensional concept, the authors, in (Abowd et al., 1999), defined certain types of context that are, in practice, more important than others: *location, identity, activity* and *time*.

  (Göker and Myrhaug, 2002) presented five main categories of context elements: *Environment* context which consists in entities that surround the user and information which is accessed by the user. *Personal* context aggregates physiological context and the mental context such as mood and expertise... The *Task* context can be described with explicit goals, tasks, actions, activities, or events. The *social* one refers to the social aspects of the current user context such as friends, co-workers, relatives and especially the role the user plays in this context.

In the *spatio-temporal* context are included aspects of the user context relating to the time and spatial extent for the user context such as time, location and direction...

From those definitions, we note that *context* is usually coupled with the user as it characterizes his situation and represents any information surrounding the entity generally and the user specifically (Fig. 3.2.

In brief, the context depends not only on the collection of the user's personal data but also on how integrating those data into a profile. In fact, the user profile represents the most important contextual element which can improve the accuracy of the search (Park, 1994).

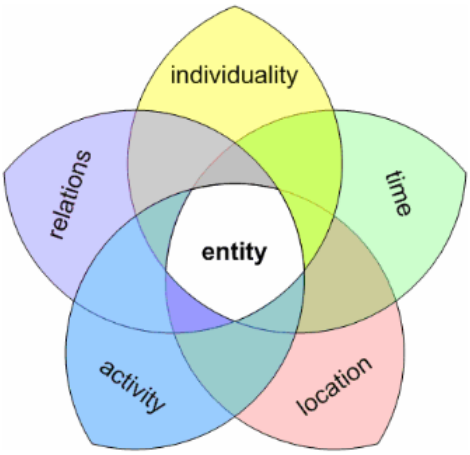

FIGURE 3.2: Five Fundamental Categories for Context Information (Zimmermann et al., 2007)

- **User Profile**

The user profile is a description of his interests and preferences allowing search systems to retrieve the most relevant results adapted to his expectations and needs.

A profile was defined by *Merriam-Webster dictionary* as : *"a brief written description that provides information about someone or something"*.

The user model was defined by (Gils and Proper, 2003) : *"A (user) profile consists of a set of preferences with regard to behavior of a search engine as well constraints on the results it presents to the user"*.

Another definition was given by (Schiaffino and Amandi, 2009): *"A user profile is a representation of information about an individual user that is essential for the (intelligent) application we are considering[...] the most common contents of user profiles: user interests; the user's knowledge, background and skills; the user's goals; user behavior; the user's interaction preferences; the user's individual characteristics; and the user's context"*

The fundamental purpose of personalized information retrieval systems (PIRS) is to tailor the user's information needs by integrating the profile in the process of information access. Its representation requires collecting

information about the user. Consequently, we need to introduce how to collect data about the user and how those data can be integrated in order to tailor search results to the user's needs and preferences.

3.3 User Profile Modeling

In this section, we present data sources that can bring out knowledge about the user and its context as well as the techniques allowing representing and building the profile.

3.3.1 Information Sources Acquisition

To personalize the Web search, we need first to get advantage from information sources that are frequently used at a daily basis by the user. The user profile can be explicitly built, by asking the user to provide his own information, or implicitly by watching the user's activities on a specific application.

- **Explicit Feeback**
  The ***Explicit Acquisition Approach*** refers to extract information provided directly from a user. It requires a user intervention to put manually information provided, generally, upon registration to Websites. Such input include demographic information, age, address... This type of profile is widely used in recommender systems, commercial Websites and personalized Websites' interfaces (Gauch et al., 2007). For instance, (MyYahoo!, 1995) (Fig. 3.3) used explicit user interaction in order to personalize the interface by choosing items to add and which content interests him. In the context of commercial Websites and recommender systems, the user is asked explicitly about his preferences in the form of ratings raged between "Highly interested" to "No interested at all".

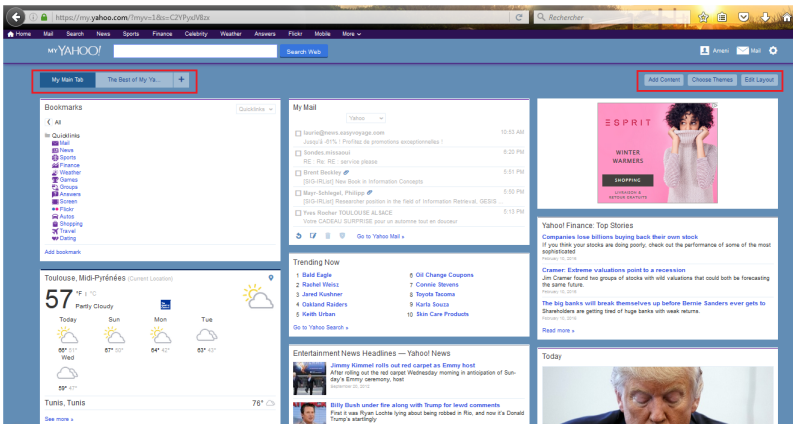

FIGURE 3.3: Sample of personalized interface (MyYahoo!, 1995)

The main limitation of this approach is that the user is usually unwilling to participate in such forms or ratings. Thus, such constructed

profile may lack of accurate information (O'Sullivan et al., 2003). In addition, as usually the user express such information in the first interaction with the system, the constructed profile is static. However, there is a need to get updated information about the user in order to improve his current search experience.

- **Implicit Feedback**
  The ***Implicit Acquisition Approach*** indicates tools allow collecting implicitly information about the user. The advantage of this approach compared to the explicit one is that there is no effort required from the user (Sugiyama et al., 2004). It is achieved by observing users' behaviors while interacting with a system such as reading time, saving, marking, purchasing and selecting (Kelly and Teevan, 2003). The types of implicit feedbacks depends on the domain:
  - Websites: visited Webpages, examination duration, ...
  - Commercial and Recommender Websites : buying, rating and recommending items...
  - Search engines: queries, Web browsing history...
  - Social Media: posts, social signals (like, share, retweet, favorite...)...

3.3.2 Information Sources for User Profiling

As the implicit feedback collect users' behaviors that allow inferring user interests, we seek to study two main implicit information sources namely *Browsing History* and *Social Media*.

- **Browsing History**
  The users' profiles are often constructed based on implicitly collected information. Hence, implicit feedback techniques take advantage of the user behavior to understand his/her interests and preferences.
  There are a lot of techniques that allow collecting browsing histories (Sugiyama et al., 2004; Morita and Shinoda, 1994; Barrett et al., 1997; Pretschner and Gauch, 1999a) containing the URLs visited by users, the dates and times of the visits as well as the time spent visiting those pages (known as Dwell Time). There are tow main approaches to collect those information:
  - *Web logs*
    (Mobasher, 2007) represent the use of a log file from a Web server, and based on the values contained in the log file, derive indicators about visits to the Web servers such as the number of visits, the number of unique visitors and visits duration... Thus, these indicators help to personalize services presented at a given Website.
  - *Search logs*
    (Sieg et al., 2004) refers to the use of data stored in transaction logs of Web search engines. Search logs can provide beneficial information about the online searchers such as the submitted

query, the corresponding search results and the URLs the user may click.

- **Social Media**
  According to *Merriam-Webster dictionary*, Social media refers to *"forms of electronic communication (as Web sites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages, and other content (as videos)"*.

  First, the main goal of most of the people who write blogs or tag a Webpage is to share their thoughts, interests or findings with others. Therefore, it is much more interesting to use social information than personal information such as emails or browsing history. Second, from the online social activities, very accurate information about users' interests can be learned. Finally, noise level, related to user preferences, is relatively low. Many online social activities, such as bookmarking and blogging, are actively initiated by users. For example, people are likely to write a blog or bookmark a Webpage about something that interests them.

  As we focus in this thesis on the IR personalization based on social systems, we lay out in this section the different social information classes presented in the literature.
  (Kaplan and Haenlein, 2010) defined social media types including tags, social bookmarking, blogs and Microblogs.

  - **Tags (Tagging)**
    With the emergence of online services, users can annotate the documents they share or create (Golder and Huberman, 2006). Such annotations may be free text comments, votes in favor of or against its quality, or tags. Tags or Folksonomies can be defined as freely chosen words assigned by Web 2.0 users in order to label content: images, videos... (Cai and Li, 2010) proposed a tag-based user profile and resource extracted from tags assigned to recipes using MovieLens sample database. (Vallet et al., 2010) introduced folksonomy-based user and document profiles. They adapted the Vector Space Model (VSM) and Okapi BM25 ranking model to social tagging profiles in order to personalize search results extracted from Yahoo! search engine.

  - **Bookmarking**
    Bookmarking is a method for Internet users to organize, store, manage and search for bookmarks of resources online. Many online bookmark management services, such as Delicious in 2003, have been launched. Indeed, bookmarks are a simple tool for building personalized subsets of information where interesting or useful Web pages (Uniform Resource Locator: URLs) can be stored for later use (Keller et al., 1997). In addition, users only bookmark documents that are relevant to them and have a motivation to add meaningful metadata to their bookmarks. (Abrams et al., 1998) discussed the usage of bookmarks and carried out a survey in order to examine users' bookmarks, their usefulness, frequency, growth and methods of organizing.

- **Blogs**
  A Blog is defined as: *"a personal journal published on the World Wide Web made of discrete entries ("posts") typically displayed in reverse chronological order so that the most recent post appears first"* (Beale, 2006). Blogs represents an opportunity to people not only to express their opinions nut also to interact and communicate with other.

  Blogs can be used by single users or by companies. Recently, blogging is an activity related to influential and inspirational personalities.

- **Micro-blogging Services** The Microblogs are services that reflect social relations among people, such as user communities and common interest groups, and allow users to share content, through typically short, but informative, text messages (Rocha et al., 1970), which may include links to images, videos or Web pages. Microblogs can be used to model the user profile (Kacem et al., 2014), to estimate the relevance of a resource (Badache and Boughanem, 2015), etc...

3.3.3 User Profile Representation

This section surveys basic techniques that allow representing and building the user profile. In particular, we describe semantic networks, concepts and vectors based profiles.

- **Vectors Profile**

  The representation of the user profile using sets of keywords has become a common way for personalization. The keywords can be explicitly provided from the user or implicitly extracted from Web documents such as visited Web pages, saved Web pages and bookmarks...

  In addition, this technique can separately treat each keyword as a topic of interest, or jointly categorize keywords into classes of user interests. Each keyword has an associated weight which represents its numerical importance in the profile (Gauch et al., 2007).

  There are three ways to represent keywords profiles: a set of weighted terms where each keyword is a single interest (Moukas and Maes, 1998; Becerra et al., 2013), a vector of weighted terms representing an interest (Lieberman, 1997; Chen and Sycara, 1998) or a set of vectors each containing weighted terms and representing a user's interest (Widyantoro et al., 1997).

- **Concepts Profile**

  The similarity between concept-based profiles and semantic network-based profiles is the fact of containing both conceptual nodes and relationships between them. However, in concept-based representation, there are brotherhood and parent-child relationships that can be easily updated compared to a semantic representation that is based on multiple relationships between words (Man et al., 2016). Generally,

this type of representation requires the use of ontologies and hierarchical concepts. Several approaches used the *Open Directory Project* (ODP)[4] (Pretschner and Gauch, 1999b; Daoud et al., 2009; Trajkova and Gauch, 2004) in order to model the user profile interests. In fact, ODP, known today as DMOZ, is a free tool that organizes Web pages in which organized content is done by volunteer editors.

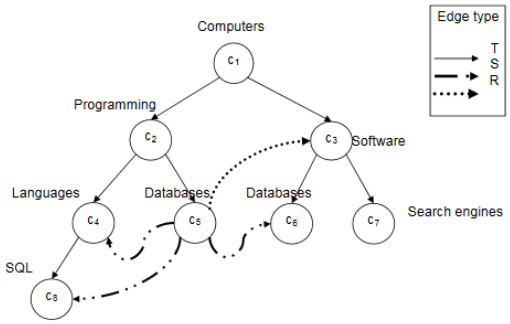

FIGURE 3.4: A Portion of an ontological user profile (Daoud et al., 2009)

- **Semantic Network Profile**
  The Semantic representation (Begg et al., 1993) is not only based on the extraction of terms from relevant documents, but also on incorporating those terms into a network of nodes. The construction of such profiles requires the creation of semantic relationships between the network nodes (Tchuente et al., 2013). In (Michlmayr and Cayzer, 2007), the authors represent proposed a tags-based user profile through a graph representation (Fig. 3.5 in which labeled nodes correspond to tags proposed by the user, and edges refers to the relationship between them.

FIGURE 3.5: A graph-based representation where Two nodes are linked with an edge if the corresponding tags have been used in combination for annotating a bookmark (Michlmayr and Cayzer, 2007)

[4] www.dmoz.org

3.3.4 Personalized Approaches to Information Retrieval

The personalization based on user profiling can be integrated using several methodologies. This section covers the primary ones identified by (Pitkow et al., 2002) : query expansion and result processing.

Query Expansion

The query expansion technique refers to: *"Query expansion refers to the process of augmenting a query from a user with other words or phrases in order to improve search effectiveness"* (Ma et al., 2007). It consists in reformulating an initial query by adding terms or by changing the weights of its terms.

A user may submit a query that contains few words and does not express exactly what he is looking for due to his little experience in web searches. To overcome this problem, the query expansion technique fits to support the user in his/her search and allow them to enlarge the search domain, to include sets of words that are linked to the frequency of the term the user specified in his/her query (Biancalana and Micarelli, 2009).

Result Processing

Another way for personalization is result processing which consists of filtering, clustering or re-ranking the list of results returned by a search engine. This technique has been used through four types (Mobasher, 2007): content-based, rules-based, and collaborative filtering.

Content-based personalization Personalized Web search can be achieved by checking content similarity between Web pages and user profiles. In content-based systems, a user profile represents the content descriptions of items in which that user has previously expressed interest (Lops et al., 2011). It resides in representing a user-vector containing interesting items to the users and their weights and predicting the items that may be of interest the user bu measuring the similarity between the user-vector and each unseen item (Mobasher, 2007).

Rules-based personalization Rules-based filtering systems depend on manually or automatically afforded decision rules that are used to recommend items to users (Romero et al., 2007). Those rules are specified according to the users' characteristics such as demographic and behavioral ones. The rues are usually defines by experts and thus may be very dependent to the domain. This approach is very used in commercial Websites in order to personalize discounts in the case of behavioral characteristics or destinations' vacations recommendation in the case of demographic features (Adomavicius and Tuzhilin, 2001). However, this type of personalization requires explicit interactions from users.

Collaborative filtering The importance of Web communities and collaborative systems has been developed with the rise of social Web such as blogs, wikis, social networks and tagging systems. There have been many researches which investigated how the search behavior of communities of like-minded users can be harnessed and shared to improve personalized search and bring more relevant search results (Jiang et al., 2015). This type of personalisation is based on two steps *(1)* detect similar users to the active user *(2)* ranking unseen items by the active user and seen by his neighborhood. (Mobasher, 2007)

In this thesis, we use the latter way of personalization, *result processing*. In particular, we use content-based personalization via the extraction of keywords from different sources and the construction of the user profile based on a vector space model. Among data sources, social media has an increasing attention in recent years as it represents a new way to the user to express his interests and connect with other people from his community.

3.4 Towards Using Socia Web in User Profiling

The Social Web has emerged in recent years and become a valuable source employed in daily basis by Internet users. In fact, it affords a new form of content that is created and shared called *User Generated Content*.

3.4.1 User Generated Content (UGC)

A lot of definitions of UGC have been proposed. In the work of (Vickery and Wunsch-Vincent, 2007), UGC was defined as *: i) content made publicly available over the Internet, ii) which reflects a "certain amount of creative effort" and iii) which is "created outside of professional routines and practices"*. Another definition was given by (Baeza-Yates, 2009): *User Generated Content (UGC) is one of the main current trends in the Web. This trend has allowed all people that can access the Internet to publish content in different media, such as text (e.g. blogs), photos or video*.

More recent work (Moens et al., 2014) defined it as: *any form of content such as blogs, wikis, discussion forums, posts, chats, tweets, podcasts, digital images, video, audio files, advertisements and other forms of media that was created by users of an online system or service, often made available via social media Websites*.

From those definitions, we notice that UGC is always related to the advent of Web 2.0 and more precisely to social media platforms. With the amount and various formats of data on that new form of Web, search systems tend to be more adapted to it in the way they create the user profile and integrate it to personalize search.

Currently, there are more than 1.6 billion social network users worldwide with more than 64 percent of them accessing social media services online (Statista, 2016). In fact, social Web is a famous way to users enabling them

to be in contact with friends and families, as well as share, interact, distract themselves and be always up-to-date.

In Fig. 3.6 , we present the most prominent social networks according to the shared content. As shown, Facebook and Twitter have the most important percentage of total sharing activity.

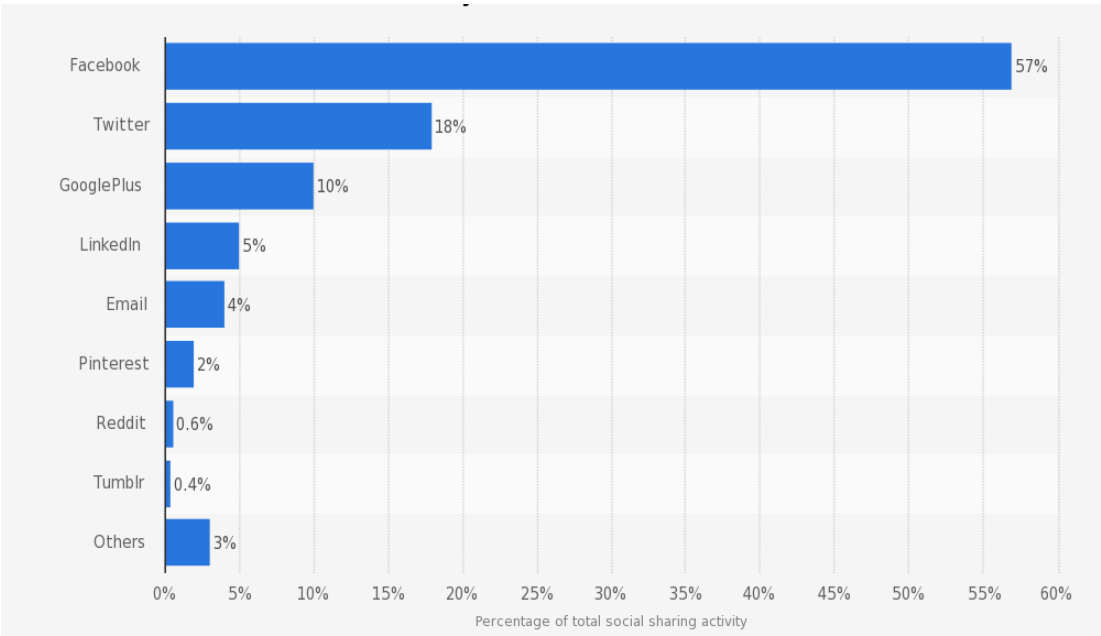

FIGURE 3.6: Distribution of global social content sharing activity as of the 2nd quarter of 2016, by social network (Statista, 2016)

3.4.2 Emergence of Social Information Retrieval

Information retrieval systems take advantages of those data that can provide implicit and explicit feedback about the user and his context. The purpose of those systems is to improve the user experience by providing relevant results to his information needs expressed on queries.

(Kirsch et al., 2006) gave a definition of Social information retrieval systems: *Social information retrieval systems are distinguished from other types of information retrieval systems by the incorporation of information about social networks and relationships into the information retrieval process*. They described a domain model for Social IR described in Fig. 3.7. They find that after moving to Web 2.0, new associations between elements appeared such as individuals that are the consumers and producers of content and queries that express individuals' information needs or knowledge. They enlightened that interactions between individuals is crucial for these systems.

Similarly the term *"Social Search"* appeared and it was defined by (Morris et al., 2010) as: *Social search may also involve conducting a search over an existing database of content previously provided by other users, such as searching over the collection of public Twitter posts, or searching through an archive of questions and answers, such as in the Answer Garden system Many works used users' annotations*. In 2013, a general definition was given by (Jeon and Rieh, 2013):

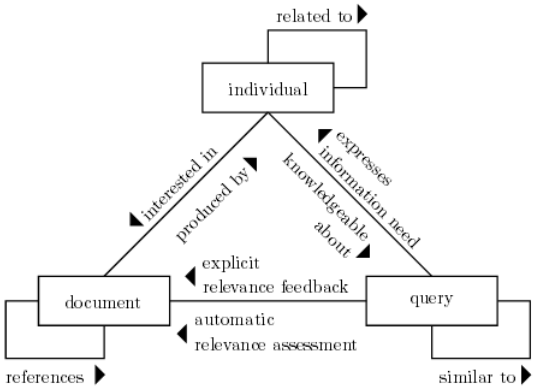

FIGURE 3.7: A domain model for social information retrieval (Kirsch et al., 2006)

Social Media Site	Foundation Date	Number of Monthly Active Users
Facebook	February 4, 2004	More than 1.59 billion
Twitter	March 21, 2006	More than 320 million
Google +	December 15, 2011	418 million
Pinterest	October 6, 2010	400 million
LinkedIn	December 14, 2002	187 million

TABLE 3.1: Popular Social Media Sites as of May 2016 (Maina, 2016)

social search refers to the process of finding information online that utilizes social resources through interactions.

3.4.3 Social-Media based User Modeling

It becomes essential for personalized search systems to integrate the social property of the Web. Data available on social Web does not give an overview of how much and how long users are present on it, but also on how to use these available data as a source to understand users' queries, discover their personal interests and preferences, and to detect their sentiments or opinions. All these are research areas used to improve users' experience with search engines and to provide them with the most relevant content at the earliest possible stage of his search.

In the table below (Table 3.1), we present the most famous social media sites in 2016 (Maina, 2016):

The social-based approaches exploit data gathered from such social systems in order to extract implicit knowledge about the searcher's preferences and interests. Current systems tend to collect information about the user by considering folksonomies as a primary source to define the user's profile since the user's annotations (Cai and Li, 2010), which are the descriptions that the user assign to describe resources on the WWW, represent his interests.

The first type of those approaches analyzes the social services in order to identify the significance of a resource. For instance, (Bao et al., 2007) adapted popularity measures for Social SimRank and Social PageRank by focusing on the folksonomy structure. The first algorithm gives the relevance of a document to a query. The second algorithm, on the other hand, measures the document popularity.

Besides, another alternative of PageRank was used for each topic, so called Topic-Sensitive PageRank (Haveliwala, 2002). Thus, pages considered important in some subject domains may not be considered important in others. In addition, (Qiu and Cho, 2006) extend the Topic-Sensitive PageRank computing multiple ranks, one for each OPD topic.

The second type of social personalization is more likely to be aware of the representation of the searcher's context. (Noll and Meinel, 2007) examined two types of profiles: the user's profile and the document's profile in order to define related tags that were used to rerank the non-personalized search results.

(Xu et al., 2008) proposed a folksonomy-based personalized search in which the user's interests and the page's topics are determined using tags extracted from Delicious and Dogear. They used this topic matching to rerank the web pages rather than using only the term matching between the query and the document.

(Carmel et al., 2009) explored the user's connections in social networks. They re-ranked search results based on their connection strength with the user's related persons and topics. In fact, they used three types of profiles: with explicit familiarity connections, with connections obtained through common social activities and finally merging both of the previous types.

According to (Paliouras, 2012), *"One of the major innovations in personalization in the last 20 years was the injection of social knowledge into the model of the user."*

Thus, the social data was introduced in the user profiling process. For instance, in (Michelson and Macskassy, 2010), the authors proposed an approach to discover users' topics of interests by examining the entities they mentioned on Twitter. They used categories to discover users' profiles chosen from candidates of each entity's page from Wikipedia.

(Garcia Esparza et al., 2013) proposed a user profiling approach described in 3.8 based on topical categorisation of user's posted URLs in order to limit the information provided on the user's timeline when starting following other users.

In (Xu et al., 2011), the authors eliminate tweets that are not related to the user's topics of interests such as tweets related to every-day life and conversations with friends. They revised the Author-Topic model (Rosen-Zvi et al., 2004) that extends Latent Dirichlet Allocation (LDA) (Blei et al., 2003) by including authorship information and introducing a latent variable to indicate whether a tweet is related to its author's interest.

User profiling based on social media and especially *Twitter* is widely used in the recommendation field. (Abel et al., 2011) introduced a twitter-based

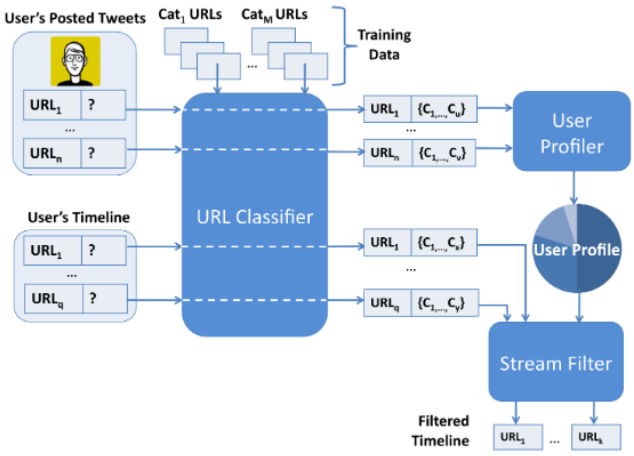

FIGURE 3.8: CatStream's System Architecture (Garcia Esparza et al., 2013)

user model for news recommendations.They analyze how strategies for constructing hashtag-based, entity-based or topic-based user profiles benefit from semantic enrichment. Their profiling strategy resides in three steps:

- Semantic Enrichment: They extracted entities and topics of tweets and used OpenCalais[5] that identifies 39 different types of entities such as persons, events, products or music groups...
- Linkage: Based on the two previous steps, they used several strategies that link tweets with external Web resources (news articles).
- User Modeling: They proposed a method generating hashtag-based, entity-based, and topic-based profiles.

In (Hannon et al., 2010), the authors evaluated a range of profiling and recommendation strategies based on Twitter for effective and efficient followee recommendation. They proposed 5 basic profiling strategies by representing users (a) by their own tweets; (b) the tweets of their followees; (c) the tweets of their followers; (d) the ids of their followees; (e) the ids of their followers, and used a simple frequency count as the term weighting function.

A large number of features was proposed by (Hong et al., 2013) based on Co-Factorization Machines (CoFM): Categorical Features, Content Features, Local Graph User Features, User Relationship Features and Temporal Features.

3.5 Conclusion

In this chapter, we were interested in covering state-of-the-art approaches and main contributions in personalized information retrieval and social information retrieval. In fact, personalized systems rely on data extracted

[5]http://www.opencalais.com

from Web but also from the social Web especially after its expansion in the recent years.

User Generated Content is a valuable source of data providing the context of a user performing search. It is a valuable source to detect implicitly the preferences and interests of a user and thus to improve his search experience.

As time represents a challenging dimension nowadays in the area of information retrieval, we introduce in the next chapter the temporal information retrieval and profiling. According to (Abel et al., 2011), it is beneficial to include a temporal factor in the user profile because content shared lately concerns the user more than content shared long time ago.

Chapter 4

Time in Information Retrieval

4.1 Introduction

In the previous chapter, we presented related work regarding standard and personalized search. We studied the integration of the user profile in the IR process. Specifically, approaches of user profile representation using both the Web and Social Media were introduced. Both areas give an overview of the context of the user.

Herein, we give attention to the temporal dimension in the IR field and more precisely in the user profile. From Fig. 4.1, we note that Internet is constantly evolving and changing over time especially with the growing amount of users each year reaching over 3,5 billion as of September 2016[1].

With this growth of Internet users and content, users are more interested to recent content and data streams. In fact, with the Web 2.0, the interactions and timelines of followers or friends influence the content shared or published by the user. The user is facing a home timeline in which his followers and friends publish content that cannot be related to his preferences but can have an impact on the user and influence his tastes and interests.

In this chapter, we study the temporal information retrieval and its main properties. By introducing the temporal information in the user profiling process, we try to answer the following problematics:

- What is Temporal Information Retrieval?

[1] www.InternetLiveStats.com

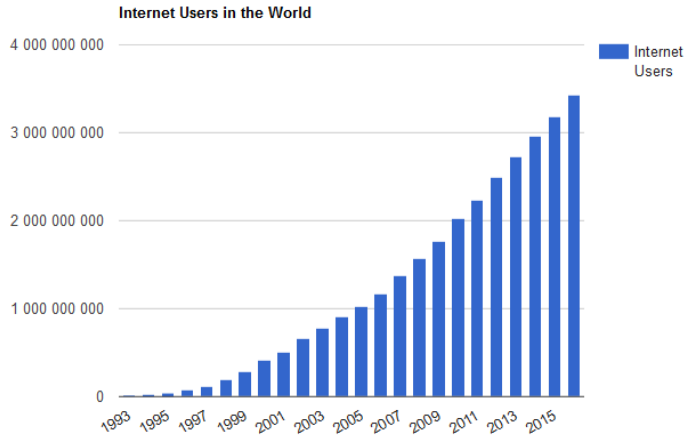

FIGURE 4.1: Internet users evolution in the World between 1993 and 2016 *(Source: www.InternetLiveStats.com)*

- How should we integrate the temporal factor in the user profile ?
- What are the benefits of modeling a temporal profile on search performance ?

To answer these issues, we investigate the temporal IR, which is the integration of time in the information retrieval, in two aspects: *(1)* Social Media *(2)* the Web especially Session Search.

4.2 Overview of Temporal Information Retrieval

Information retrieval aims at satisfying a user's need expressed in a form of a query with relevant documents. As time has gained increasing importance in recent years, a new research area has appeared known as *Temporal Information Retrieval (Temporal IR)*.

Time can be defined by The American Heritage dictionary as: *"A nonspatial continuum in which events occur in apparently irreversible succession from the past through the present to the future"*. It was formally defined by (Bruce, 1972) as *"An ordered pair, (time,$\leq$), where time is a set whose elements are called time points and $\leq$ is a relation that partially orders time."*.

By integrating time in the information retrieval filed, (Campos et al., 2014) gave a definition of Temporal IR: *"In general, T-IR aims to satisfy search needs by combining the traditional notion of document relevance with temporal relevance."*

The temporal dynamics was first introduced to study the change of the Web, its forms and content. In fact, Web has been changed in different ways: volume, content, structure, users' behaviors and needs... Detecting the changes on users' behaviors is one key aspect that determine interests' relevance to each user.

In (Alonso et al., 2011), the authors gave 11 research trends in the Temporal IR field namely: Exploratory Search, Micro-blogging and Real-time Search, Temporal Summaries, Temporal Clustering, Temporal Querying, Temporal

Question Answering, Temporal Similarity, Timelines and User Interfaces, Web Archiving and Spatio-temporal Information Exploration.

In (Moulahi et al., 2015), the authors classified temporal approaches into three categories: *Query Level* where works attempt to understand the period specified in queries and the temporal intent behind them, *Document Content Level* where studies extract and represent temporal expressions contained in documents, and at *Document Matching Level* where time is integrated in the ranking process. (Campos et al., 2014) provided more fined categorization through a variety of tasks: Web Crawling and Web Archiving, Indexing, Query Processing, Temporal Ranking, Temporal Clustering, Temporal Text Classification, Temporal Search Engines, Future-Related Information Retrieval...

Temporal IR is present at two main aspects according to (Kanhabua et al., 2015): *(1)* Content and structure changes *(2)* User behavior changes.

Content and Structures Changes As the content of the Web is constantly changing over time (Web pages modification, addition, removal), there were initiatives that allow achiving this content. The most known one is *Internet Archive*[2] which is a non-profit digital library that was launched in 1996 by Brewster Kahl that has collected over 505 billion pages (as of September 19th, 2016). The authors proposed a categorization of documents changing over time such as, personal homepages, corporate websites, Wikipedia articles and blogs... They used two types of categories: static or dynamic.

Changes in User Behavior This type of change is related to the user attitude when interacting with search engine or Websites. In fact, queries issued by the user may be influenced by the time of its submission (weekday, weekend...) or public trends. Moreover, queries may be classified as time-sensitive queries when they include temporal expressions or an implicit temporal information.

In our thesis scope, we are interested in the second category. In fact, we aim at modeling the user profile considering the temporal distribution of his interests.

According to state-of-the-art approaches the evolution of a constructed user profile resides in its adaptation to changes in user interests over time. It is often done by an incremental process based on the addition of new information in the representation of the profile. The user profile evolution consists mainly of capturing interests' changes and update the profile content accordingly. Furthermore, prior work discern the short-term and long-term profiles. The first one represents interests related to the user's current search activities. The second one represents user's persisting interests which are extracted from his entire search history.

Short-term and long term in addition to other temporal approaches regarding user profile evolution are summarized in Fig. 4.2. We present in the following sections these approaches in order to study the evolution of the

[2] www.archive.org

user profile in both contexts *(1)* Web search, particularly in Session Search, *(2)* Social media.

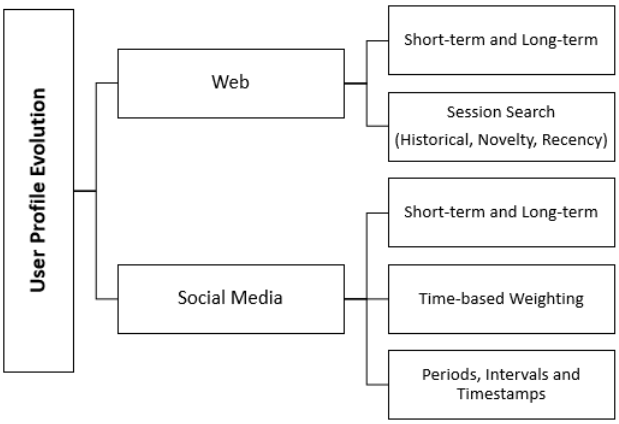

FIGURE 4.2: Summary of User Profile Evolution Approaches in both Web and Social Media fields

4.3 Temporal User profile on the Web

In this section, we study the state-of-the-art approaches introducing the temporal information into the user profile and the retrieval process. We present, thus, approaches addressing short-term and long-term profiles used both and separately. Then, we focus on session search as time takes action in the interactions' sequence.

4.3.1 Web-based Short- and Long-term Profiles

As the Web content is continuously evolving in terms of amount and formats, it becomes necessarily for users to go through a lot of irrelevant documents until they select the relevant ones. This becomes an overwhelming process that requires the adaptation of Web documents to the needs and preferences of each specific user obtained from his profile and knowledge extracted from it. In order to track changes in the user profile and to predict better his needs and expectations, most of the prior work discern two types of profiles: *the short-term and long-term*.

Web-based Short-Term Profile The short-term profile describes the interests and needs of users related to current activities and search session. The context of search activities within the current session has been used to build richer models of interests and improve how the search system interprets the user's current query.

It is assumed that the exploitation of short-term profile, as the immediately related preceding query and the viewed documents in the same session, is

used to target better the search as it contains specific information considered relevant to the immediate information needs of the user (Shen et al., 2005).

The main goal of profile's changing in the short term is to improve the accuracy of information retrieval using the most appropriate profile without noise caused by the interests which are not related to the search context. Therefore, this profile allows efficiently adapting the IR process to the specific information needs of the user.

Some approaches (Dumais et al., 2003; Pretschner and Gauch, 1999a) do not necessarily consider the same information need but can reflect multiple interests. In these approaches, the evolution of the user profile is related to the delimitation of the recent activities of the user by a time interval which may include several interests. Other studies (Tamine-Lechani et al., 2008) define the short-term user profile in a search session by a single information need. The evolution of the profile in this case requires search sessions' boundaries mechanisms, where a session is defined by a set of queries related to the same information need. (Daoud et al., 2009) represent a short term user interests based on the user context in a particular search session via a set of weighted concepts. It is built and updated across related queries using a session boundary identification method.

Besides, In (Shen et al., 2005), the authors propose to improve how the search system interprets the user's current query using short-term profiles based on search queries and result clicks.

(Xiang et al., 2010) developed heuristics to promote search results with the same topical category if successive queries in a search session were related by general similarity, and were not specializations, generalizations or reformulations. In the work of (White et al., 2010), the authors aim at predicting short-term interests by modeling the search context of a user. They studied the intent of a search by using different sources: *Query*, *SERPClick* and *NavTrail* corresponding to ODP labels assigned automatically to top-ten search results, those clicked by the user and Web pages that the user visits after a SERP Click.

To address the problem of short-term personalization, (Ustinovskiy and Serdyukov, 2013) employed short-term browsing history and used a variety of features related to the query, the browsing session as the context of the query, click-through-based proximity and SERP-aggregated features.

Web-based Long-term Profile The other kind of profile is long-term profile, which refers to the use of specific information such as the user's education level and general interests, occurred user query history and past user click-through information. In fact, such information are generally stable for a long time and are often accumulated over time. The long-term profile can be applicable to all sessions, but may not be as effective as the short-term profile in improving search accuracy for a particular session.

Filtering systems, such as Grouplens (Konstan et al., 1997), are among the first systems used to update long-term profile. (Teevan et al., 2005) developed rich long-term user models based on desktop search activities to

improve ranking. Besides, (Matthijs and Radlinski, 2011) developed models of users' interests using a combination of content and previously visited Websites. The evaluation is based on an interleaving methodology merging original and personalized rankings. (Tan et al., 2006) studied long-term language model-based representations of users' interests based on queries, documents and clicks. They considered different amounts of history and found that for fresh queries recent history was the most important, but for recurring queries longer-term history was more significant. Recently, (Sontag et al., 2012) developed generative and discriminative probabilistic models using ODP category from historical click data. The parameters of the particular user in this generative model constitute a condensed user profile learned from a user's long-term search history.

Usage of Short- and Long-term Profiles on the Web In (Bennett et al., 2012), the authors proposed a novel unified framework in order to study the dynamics of user behavior. They personalized an issued query using both short- and long-term user profiles based on three temporal views as described i Fig. 4.3: session, historical and aggregated. They considered issued queries and previous results and any action performed by the user such as viewing a result, explicitly ignoring it or missing it.

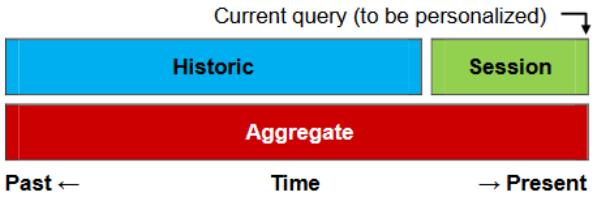

FIGURE 4.3: Temporal views of user profile: recent (Session), past (Historic), or a combination (Aggregate) (Bennett et al., 2012)

(Tamine-Lechani et al., 2008) proposed a personalized document ranking based on a user profile. First, they modeled the user profile that contains concepts of interests inferred from his browsing history. Second, they learned the user long-term interests by managing the short-term interests. They defined the short-term profile as a limited number of sessions and use it to update the profile using a correlation measure allowing detecting changes in user behavior.

In the same scope, (Li et al., 2007) studied learning user profiles and used them to re-rank search results. They assume that long-term profile contain stable interests during a long period while the short-term profile is unstable and change over time. For the long-term profile, the authors used Google Directory and extracted interests from Web search results and linked as a tree structure with a preference score as shown in Fig. 4.4. For the sort-term profile, the authors frame the Page-History Buffer (PHB) that detects the most recently clicked pages with a fixed size according to the ability of the search engine.

In the table 4.1, we present a summary of the previously discussed approaches based on the profiles' type.

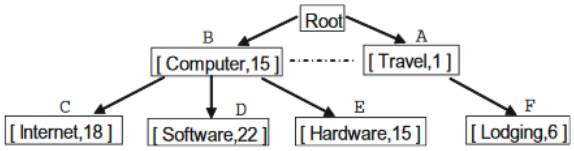

FIGURE 4.4: Model of Long-Term User Profile (Li et al., 2007)

Profile Type	Related Work	Profile Data Source
Short-Term Profile	(Shen et al., 2005)	Previous queries, Results
	(Dumais et al., 2003)	Desktop, Web
	(Daoud et al., 2009)	Annotated queries, top n relevant documents
	(Xiang et al., 2010)	Search log from a commercial search engine
	(White et al., 2010)	Logs of visited URLs
	(Ustinovskiy and Serdyukov, 2013)	Commercial search engine data
Long-Term Profile	(Konstan et al., 1997)	Usenet System News
	(Teevan et al., 2005)	Web search results
	(Matthijs and Radlinski, 2011)	Queries, Clicked results
	(Sontag et al., 2012)	Bing search engine results
Short- and Long-term Profiles	(Tamine-Lechani et al., 2008)	News documents
	(Bennett et al., 2012)	Microsoft Bing search engine logs
	(Li et al., 2007)	Google search results

TABLE 4.1: Summary of Temporal User Profiling Main Approaches

4.3.2 Session Search in Focus

Search history is known to be a valuable source of information about user's preferences and search purpose. Time is integrated to study the queries sequence and their impact on current search task. Session search is defined by (Luo et al., 2015) as: *"an information retrieval task that involves a sequence of queries for a complex information need. It is characterized by rich user-system interactions and temporal dependency between queries and between consecutive user behaviors"*. From this definition, we can conclude that the main goal of session search is to improve a current query results taking into consideration user behavior during a session. In fact, a session comprises: a current query, to which we need to predict results, a set of previous interactions containing past submitted queries and clicked results.
We group the approaches addressing session search into three categories *(1)* approaches analyzing the user behavior through historical queries, *(2)* those integrating the novelty *(3)* employing recency of results. Each of these approaches is detailed below even though only the last category is related to our thesis scope but the two first types of approaches are significant to personalize session search.

Historical Queries within Session Search

We present the usefulness of historical queries that resides in exploring relationships between current and previous queries in terms of terms change and reformulation. This category is mainly based on query expansion (Section 3.3.4).

In (Zhang and Yang, 2013), the authors proposed an approach for current query change using the Markov Decision Process (MDP) by decomposing each adjacent query-pair into three parts: the added terms, the removed terms and the theme terms (common terms between queries) based on cross-session information. Similarly, (Guan et al., 2013) proposed a novel query change retrieval model (QCM) based on MDP too. To enhance session search, they utilized syntactic changes between nearby queries and the relationship between query change and previously retrieved results.

(Chen et al., 2013) used query expansion based on historical queries and the current query. They used unigram, bigram, 3-gram and 4-gram phrases to detect the entity candidates and then weighted each term or phrase accordingly. In (Matthias et al., 2013), the authors exploited conservative query expansion strategy based on past queries/ clicked documents, similar sessions from other users and their clicked results. Specifically, they used different strategies of segmenting the queries for identifying and underlining concepts in the queries.

In (Liu et al., 2010), the authors studied the influence of both task type and situation on user's query reformulation behavior. A taxonomy of query reformulation was proposed based on five reformulation types: Generalization, Specialization, Word Substitution, Repeat and New. In the table below, we present the five reformulation types illustrated by examples considering Q_{i+1} is the query following the query Q_i in a same session. They evaluated

Generalization (G)	Q_i and Q_{i+1} contain at least one term in common; Q_{i+1} contains fewer terms than Q_i	"harmful chemicals in food" → "chemicals in food"
Specialization (S)	Q_i and Q_{i+1} contain at least one term in common; Q_i contains more terms than Q_i	"2007 car" → "2007 car sales"
Word Substitution (WD)	Q_i and Q_{i+1} contain at least one term in common; Q_i has the same length as Q_i, but contains some terms that are not in Q_i	"castle in canada" → "fortress in canada"
Repeat (R)	Q_i and Q_{i+1} contain exactly the same terms, but the format of these terms may be different	"Danmark fortress" → "fortress, danmark"
New (N)	Q_i and Q_{i+1} do not contain any common terms	"anthill" → "ant bites"

TABLE 4.2: Description of five query reformulation types illustrated with examples (Liu et al., 2010)

these categories using three task behaviors: *(1) Simple task*: where the information sought is single and independent *(2) Hierarchical task*: where a single topic is explored through different facets *(3) Parallel task*: where different concepts are explored at a same level in a hierarchy.

Novelty within Session Search

Novelty was defined by (Li and Croft, 2005): *"Novelty or new information means new answers to the potential questions representing a user's request or information need"*. It resides in identifying new information from an entity, a sentence or or document.

In (Jiang and He, 2013), the authors considered that there is a need to explore approaches balancing performance and novelty. They assumed that the user's interest degrades each time viewing a result, and proved the usefulness of past queries in whole-session search performance and the effectiveness of click-through information on maintaining the novelty. In a previous work of the same authors (Jiang et al., 2012), a method to eliminate duplicate results in ranking was introduced. They simulated users' browsing behavior in a session search under the assumption that a user reformulates a query to find not only relevant but also novel results.

Novelty and *Diversity* are related in many prior works. However, in (Vargas and Castells, 2011), the authors explain the difference between the two concepts. In fact, novelty is the new information that is different of what a user or group of users have seen or checked. Diversity refers to the difference between items and the fact of being diverse and varied. They used a similarity-based item novelty as follows:

$$nov(i \mid \theta) = \sum_{j \in \theta} p(j \mid choose, \theta, i) d(i, j) \tag{4.1}$$

where :
$p(j \mid choose, \theta, i)$ is the probability that a user choose an item *j* in the context θ after having chosen item *i*,
$d(i, j)$ is the complement of the distance between items i an j : $d(i, j) = 1 - sim(i, j)$

In order to satisfy the user intent, (Gollapudi and Sharma, 2009) used an axiomatic framework using the measures of relevance and novelty. The novelty is obtained by computing the number of categories represented in the list of top-N results for a given query. In addition, (Clarke et al., 2008) presented a framework for evaluation that emphasizes novelty and diversity based on the probability ranking principle (PRP). They stated the principle as follows: *"If an IR system's response to each query is a ranking of documents in order of decreasing probability of relevance, the overall effectiveness of the system to its user will be maximized"*.

Recency and Time within Session Search

Using session search, relevant information can be operated in different ways but they all have the same aim: delivering the most relevant results to the current information need. In line with our thesis scope, considering the temporal sequence of new additional information to the user profile during a session represents one way to insure search performance. This type of approaches is used for time-aware ranking such (e.g. news and events). It is based on enhancing recently seen or updated content (Kanhabua and Anand, 2016).

In (Baskaya et al., 2012), the authors addressed the analysis of session strategies effectiveness over time. They proved thanks to the time-based evaluation that the more time is available the less it matters how a user searches. (Kotov et al., 2011) modeled a classification framework based on features of individual queries and long-term user search behavior at different granularity. Their contribution had an impact on complex information needs and on cross-session search tasks.

Another approach enhancing recent content was proposed by (Jatowt et al., 2011) based on evaluating the degree to which search results contain fresh information and the *focus time* of Web pages. They applied the clustering to detect events and measured the freshness as follows:

$$F_p^{time} = \frac{1}{k} \sum_{i=1}^{k} w_l * sim(v_p, v_l^{event}) \tag{4.2}$$

where v_p is a feature vector of page p, v_l^{event} is the centroid vector for each cluster w_l denotes the value of the cluster l measured as follows:

$$w_l = \frac{csize(l)}{\max\limits_{1 \leq i \leq k}} . e^{\lambda . \frac{t_l - t_{beg}}{t_{end} - t_{beg}}} \tag{4.3}$$

where $csize(l)$ is the size of the cluster l (the number of its documents), t_l is the event's occurrence time and λ is a parameter controlling the influence of cluster age.

Moreover, in (Inagaki et al., 2010), the authors used a set of session-based click features to improve machine learned recency ranking. They proposed a time weighted click through rate (CRT) as follows:

$$CRT(query, url) = \frac{\sum_{i=1}^{T} c_i}{\sum_{i=1}^{T} v_i} \tag{4.4}$$

where c_i and v_i are respectively the number of clicks and views on day i during the period T. In order to enhance recent observations, the authors proposed a time varying CRT where clicks and views counts for a day are exponentially weighted:

$$CRT^w(query, url, t_{query}) = \frac{\sum_{i=1,ci>0}^{t_{query}} (c_i.(1+x)^{i-t_{query}})}{\sum_{i=1,vi>0}^{t_{query}} (v_i.(1+x)^{i-t_{query}})} \tag{4.5}$$

where x is a positive variable indicating the steepness of the recent observations' enhancement, and t_query is the time at which the query is issued

In (Diaz, 2009), the authors proposed a approach to estimate the "newsworthiness" of a query. They used clicks from different queries with the same context under the assumption that they will provide significant evidence. They used thus the Beta prior over p_q^t using a click probability π_q^t:

$$\widetilde{p}_q^t = \frac{C_q^t + \mu\pi_q^t}{V_q^t + \mu} \tag{4.6}$$

where:
$\widetilde{p}_q^t$: the predicted probability of a user click,
C_q^t: clicks observed for query q for views before t,
V_q^t: items (views) presented for query q before t,
μ : the hyper-parameter of the model.

In (Neubauer et al., 2007), the authors proposed an algorithm to improve the query's original ranking by using the one trained on feedback on the nearest query which is chosen based on different distance measures. They proved the performance of the standard ranking's results fusion with those returned by rerankers. We summarize in Table 4.3 the metrics used in their work:

Measure	Explanation and Formula
Random	It assigns an arbitrary distance: $d_{Random}(q_1, q_2) = randon(0, 1)$
OnlyQuery	It computes the topical similarity using the cosine between the term vectors, where *bin* is a function creating a binary term vector of queries: $d_{OnlyQuery}(q_1, q_2) = 1 - cos(bin(q_1), bin(q_2))$
TopN	It uses the first top N results of a query that are added to its original term vector: $d_{TopN}(q_1, q_2) = 1 - cos(TopNExpand_N(q_1), TopNExpand_N(q_2))$
Common Relevant	It is based on the ratio of shared relevant documents, where $rel(q) \subset D$ is the set of relevant documents d for a query q: $d_{CommonRelevant}(q_1, q_2) = 1 - \frac{\lvert rel(q_1) \bigcap rel(q_2) \rvert}{\lvert rel(q_1) \bigcup rel(q_2) \rvert}$

TABLE 4.3: Summary of distance measures used in (Neubauer et al., 2007)

From previous works, we note that a temporal function is highly recommended to be integrated in order to track additional information and give more importance to the most recent evidence in the user profile.

4.4 Temporal User profile on Social Media

User profiling in social media is evolving faster than any other support. In fact, on such Websites, the user shares his opinions and interests and is also faced with his community's preferences and published posts.

More precisely, micro-blogging services have been widely used to detect users' topics of interests and preferences. Users express, share and mark as favorite the content that interests them with a 140-character post (*Tweet*).

We observed state-of-the-art approaches regarding temporal dynamics in user profiling and the integration of time to improve the search experience, and classified prior works into three categories: *Short- and Long-term Profiles*, *Time-based weighting* and *Periods, Intervals and Timestamps*.

4.4.1 Social Media-based Short-term and Long-term Profiles

This category contains approaches that separate recent and old interests. Researchers assume that recent content shared or published by the user should be integrated to the short-term profile whereas long-term profile should contain recurrent interests and personalization depends on the information need relation to the recent or persistent profiles.

In (Li et al., 2014), the authors proposed a novel recommendation approach, in which the long-term and short-term reading preferences of users are regularly integrated when providing news items.

They first create a hierarchy from news articles. Then, they form news groups that correspond to the user preferences based on the long-term profile. The short-term profile was used to select news items from each selected news groups. They used an exponential temporal function to enhance temporal dynamics in the news-based user profile: $f(t) = e^{-\lambda t}$.

(Tchuente et al., 2010) focused on temporal graphs' visualization of users' interests. From a case study on Facebook, they used an approach similar to the construction of the users' semantic profiles (Gauch et al., 2007), but with two major advantages: (a) visualization of the evolution of each user interests in the form of a graph allowing to detect the user's short-term and long-term interests, (b) visualization of the influence of the social connections on user's interests by dynamic graph generation.

Besides, (Billsus and Pazzani, 2000) proposed an adaptive news access framework based on machine learning algorithm designed to extract user models. They used both explicit and implicit user feedback. They used short-term profile obtained through Nearest Neighbor Algorithm containing information about recently rated events. They converted news stories to *TF.IDF* vectors and used the cosine similarity measure to quantify the similarity of two vectors. The long-term profile was identified using a Naïve Bayesian Classifier that models a user's general preferences for news stories and computes predictions for stories that could not be classified by the short-term model.

4.4.2 Time-based Weighting

In this category, authors used a temporal-based weight to assign a temporal value to profile content. Usually, prior words were based on forgetting mechanisms and decay functions.

(Li et al., 2012) used Latent Dirichlet Allocation to extract topics from the user's messages. They clustered them to represent the user's interests with the centers of the clusters. They introduced a time-varying function to detect the influence of objects on the clustering process.

In social media field, the analysis of time within user profile has also been proposed to recommend relevant content to user. To personalize recommendation, (Abel et al., 2011) explored the temporal dynamics on the microblogging netwrok Twitter. They created two types of profiles: hashtag-based and entity-based. In both of them they integrated the time to detect profile topics of interest. They used time-sensitive variant which alleviates the occurrence frequency according to the temporal distance between the concept occurrence time and the given time-stamp:

$$w(c, time, T_{tweets,u}) = \sum_{t \in T_{tweets,u,c}} (1 - \frac{|time - time(t)|}{max_{time} - min_{time}})^d \qquad (4.7)$$

where $T_{tweets,u,c}$ is the set of tweets published by user u mentioning topic c, $time(t)$ represents the timestamp of a tweet. max_{time} and min_{time} denote the highest and lowest timestamps of tweets, and d is parameter used to adjust the temporal distance.

(Yan et al., 2012) proposed a human dynamic model co-driven by interest and social identity. They used the microblogs' messages behavior, comments and forwards in the Mobile context and supposed a linear decline of interests. They proved that as the time goes by, the user interest decreases regularly. Under this assumption, many works integrated an exponential temporal function $f(t) = e^{-\lambda t}$ in order to have a decay rate λ that reduces the weight of interests (Li et al., 2014; Ding and Li, 2005). Similarly, (Orlandi et al., 2012) proposed to aggregate distributed users profiles extracted from different social Websites in order to obtain a more complete picture of the user's profile. The authors used Facebook and Twitter accounts to leverage user interests and proposed the following decay function:

$$x(t) = x_0.e^{-\frac{t}{\tau}} \tag{4.8}$$

where $x(t)$ is the amount at time t, $x_0 = x(0)$ is the initial amount, $\tau = \frac{1}{\lambda}$ is a constant called *mean lifetime* and λ is a positive number called the *decay constant*. The time forgetting function was integrated into topic-based user interest profiling in (Tang et al., 2013) as follows:

$$\tau = \left(\frac{D-t}{D}\right)^{\mu} \tag{4.9}$$

where D is the total number of days indicating the whole time range of the dataset used by the authors. t is the count of days from the user entry date to the last date of the whole dataset. $\mu \in (0, +\infty)$ is a parameter for adjusting the forgetting rate.

4.4.3 Periods, Intervals and Timestamps

In the third category authors slot the time into intervals, periods or stamps in order to personalize or recommend content to users.
In (Yin et al., 2014), the authors analyzed user behaviors in social media systems based on *temporal context-aware mixture model* (TCAM), a latent class statistical mixture model. They proved that the user's behavior is generally influenced by intrinsic interest as well as the temporal context (e.g., a public trend). They conducted experiments on four real data sets: Digg, Movielens, Douban and Delicious.
(Jain et al., 2013) investigate temporal aspects for user behavior in Twitter also. In order to observe the evolution of a topic, they used a list of tweets on the topic at any given time and a weight to measure its strength. They studied the role of frequent users to keep the network alive and to enhance the popularity of topics.

A probabilistic framework was proposed by (Ramasamy et al., 2013) for mining user interests from their tweet times and the timing of external events associated with these interests. In Fig. 4.5 the tweet times of the user are marked by arrows. Event times are marked in red and all other

times are non-event times. In the top, the tweeting behavior is the one of a person not interested in X, whereas in the bottom we find the behavior of a person interested in X.

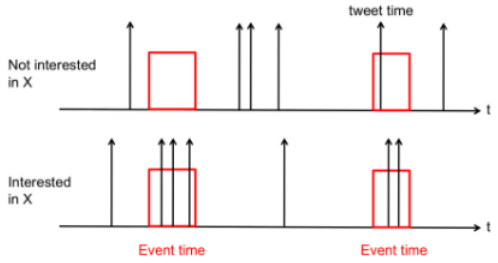

FIGURE 4.5: User Interests Inference Described by an Example (Ramasamy et al., 2013)

In (Jabeur et al., 2012), the authors estimated the tweet relevance based on the microblogger influence and the time magnitude. They observed the amount of retweets during a same period o_e of a tweet publishing. They estimated the probability $P(o_e \mid \vec{k})$ and weighted the periods as follows:

$$w_{o_e,\vec{k}} = \frac{log(\theta_q - \theta_{o_e})}{log(\theta_q - \theta_{o_s})} \times \frac{df_{\vec{k},o_e}}{df_{\vec{k}}} \tag{4.10}$$

where:
θ_q, θ_{o_e} and θ_{o_s} correspond to timestamps of the query q, the period o_e of term publishing and the period o_s when the term configuration $\vec{k}$ was firstly used (oldest timestamp).
$df_{\vec{k},o_e}$ is the number of tweets published in o_e mentioning the term $\vec{k}$, and $df_{\vec{k}}$ is the total number of tweets containing the term configuration $\vec{k}$.

In addition, (Bizid et al., 2015) proposed a temporal sequence representation in order to detect prominent users during events. They used a various of features during a timestamp based on user activities regarding an event *(on-topic activity)* and other topics *(off-topic activity)* as illustrated in Fig. 4.6. This representation is based on a Mixture of Gaussian Hidden Markov Model (MoG-HMM) where $V_u^{(t_i)}$ is the temporal sequence of user activities in the form of a set of concatenated feature vectors computed at all the timestamps t_i , S represent hidden states, A is the state transition probability matrix to change from state S_i to S_j.

4.5 Limits and Research Questions Awarded in this Thesis

In this chapter, we were interested in covering state-of-the-art approaches and main contributions in the Temporal Information Retrieval field. Research in this field has gained a lot of attention recently and there are many questions that require to be further investigated.

From the analysis of previous work, we find that the fact of discerning the short-term and long-term user profiles does not necessarily reflect the

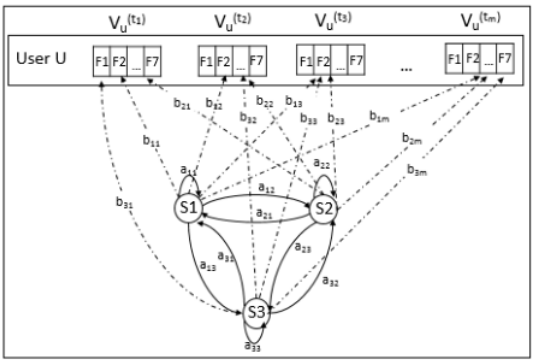

FIGURE 4.6: Temporal representation of user behavior during an event using a 3-state ergodic HMM (Bizid et al., 2015)

user's needs. For users who are not very active on social services, the short-term profile can eliminate relevant results which are more related to their personal interests. This is because their social activities are few and separated over time.

In addition, for users who are very active, the aggregation of recent activities without ignoring the old interests would be very interesting because this kind of profile is usually changing over time. In fact, considering recent interactions in this context can eliminate relevant information about users' intended purpose of search.

Furthermore, most of the approaches proposed on social media rely on time-window or time-stamps in order to track the evolution of users interests or also to capture the changes on the network such as evolution of topics/trends or the emergence of events.

To answer these challenges, we address the following research questions:

- Which method can we propose in order to track the changing of the user's interests?
- How can our model integrate the freshness of the user's interests, preferences and goals in order to merge both the short-term and long-term interests?

Through the contributions that we presented in the next chapters, we will try to answer those problematics assuming that a user profile can reflect both the recurrent (persistent) and the current (recent) interests but with different scales based on freshness.

Part II

On Using Time-Sensitive User Profiling for Personalized Search

Chapter 5

Time-Sensitive User Profile

5.1 Introduction

In order to answer the user's information need, IR systems used to match the query and the documents. They analyze the distribution of the query's terms in the documents' content. However, the information need expressed in query's terms could reflect multiple meanings or could not describe the exact meaning the user desires to characterize.

In this chapter, we present our user profiling approach based on his activities and actions on social-media Websites. We combine the frequency of those actions with a temporal weight giving more value to recent ones. Considering the temporal dynamics of the profile is one of the challenging issues in nowadays research works. The users' interests are often evolving and changing over time and may be influenced from time to time.

We evaluate our time-sensitive user profile in the context of social media. More precisely our user profile data source is the microblogging system Twitter[1]. In fact, the popularity of the Social Web makes it as an invaluable

[1]https://www.twitter.com

source of data that can help personalized IR systems achieve the goal of adapting search results to users. We note that among 3.715 billion Internet users, 2,206 billion use social networks every month[2]. For the most popular social networks, Facebook is ranked in the first place and Twitter in the fourth one [2].

5.2 Problem Description and Objectives

Many personalized search approaches explore user's social Web interactions to extract his preferences and interests, and use them to model his profile. Users interact with each other by creating and sharing content and by expressing their interests on different social Websites (Orlandi et al., 2012).

We aim at using those data in order to build a time-sensitive user profile that describes users' interests and helps a better adaptation of search results. This profile is very useful to track user's feedback regarding available content. (Mostafa, 2005) specified that:

"New search engines are improving the quality of results by delving deeper into the storehouse of materials available online, by sorting and presenting those results better, and by tracking your long-term interests so that they can refine their handling of new information requests. In the future, search engines will broaden content horizons as well, doing more than simply processing keyword queries typed into a text box."

So, search engines need to have deep knowledge about the user's activities and not be limited to query's keywords in order to improve the search experience and to understand the user's intent.

In summary, we address the following research challenges and questions:

- **How can accurate information about the user's preferences and interests be collected and represented without user involvement?**
  Detecting users' interests is an essential step to model a profile useful for personalization.
  Users are often not willing to fill in online forms in order to describe explicitly their areas of interests. Thus, we propose to implicitly extract users' activities in order to detect relevant information indicating his interests and preferences.

- **Can time-sensitivity be integrated into the user profiling strategy and be source of evidence for personalization?**
  Many of state-of-the-art approaches assign more importance to the frequent terms no matter the moment of use. Time is often used to discern short-term and long-term user profiles. Short-term profile considers current actions while the long-term profile is built according to several previous actions.

  As we already discussed in Chapter 4, discerning the short- and long-term interests requires the use of a time interval that may include

[2]http://www.blogdumoderateur.com/50-chiffres-medias-sociaux-2016/

several interests (Dumais et al., 2003), or session's boundaries mechanisms where a session is defined by a set of queries related to the same information need (Shen and Zhai, 2003; Daoud et al., 2009).

In this thesis, we propose a novel user model integrating the freshness of the user's interests, preferences and goals. Specifically, we study leveraging user's activities for user modeling and evaluate the impact of temporal dynamics on enhancing the quality of user models in the context of personalized search.

5.3 Time-Sensitive Profiling Approach

We propose to construct the user profile implicitly from his social Web activities and represent it as a vector of weighted terms which correspond to the user's interests.

In the classical non-time sensitive approaches, the relevance of an interest in the user profile is assumed to be only decided by the counts of terms in the profile, but not by their position in time. As the user interests evolve, we aim at weighting the profile's terms according to both the freshness and the frequency in order to unify both the recent and persistent interests instead of using the delimitation of session activities.

5.3.1 User profile Model

In our work, we use bag-of-words model in which extracted content is represented as a list of keywords. We collected keywords from the user interactions and compute their weights by combining both their frequency and their appearing moment. This model is widely used in document classification, image recognition (Jegou et al., 2008; Philbin et al., 2008) and visual search (Sivic and Zisserman, 2008), as well as in the information retrieval filed.

More formally, we consider a document $D^{S_j} = (t_1, t_2...t_N)$ generated at moment S_j (day, hour or minute...). In our work, by document we mean a content generated by the user such as a tag or a tweet.

In each date S_j, we define the user profile as a vector $\overrightarrow{U}$ of terms and their corresponding global weights W (Equation 5.1). We used the *Vector Space Model* that was first introduced by (Salton, 1971). It is based on representing documents as vectors of weighted keywords. The weight measures how important the term is and how effectively it reflects the document content.

We assigned a time-sensitive weight to words that reflects how relevant the term (interest) is to be for the profile.

$$\overrightarrow{U} = ({t_1}^{S_j} : {W_1}^{S_j}, {t_2}^{S_j} : {W_2}^{S_j}, ..., {t_m}^{S_j} : {W_m}^{S_j}) \tag{5.1}$$

where the temporal weight $W(t)^{S_c}$ of a term t in the profile is the sum of its time-biased relative frequency defined as follows:

$$W(t_k)^{S_C} = \sum nTF(t_k)^{S_j}.K(S_C, S_j) \tag{5.2}$$

In fact, we extract documents' terms and generate their normalized term frequency (nTF) described in Equation 5.3 after applying the text processing steps explained in Section 2.3: *Indexing*, *Stop-words removal* and *Stemming*.

$$nTF(t_i)^{Sj} = \frac{freq^{S_j}(t_i)}{\sum_{\forall k \in D^{S_j}} freq^{S_j}(t_k)} \tag{5.3}$$

with:

$freq^{S_j}(t_i)$ is the frequency of a term t_i in D^{S_j}
and $\sum_{\forall k \in D^{S_j}} freq^{S_i}(t_k)$ represents the sum of the frequencies of all terms appeared in $D^(S_j)$.

Many works have introduced the concept of *Freshness* (Badache and Boughanem, 2015; Bambia and Faiz, 2015) defined by (Bouzeghoub, 2004) as "*the most important data quality attributes in information systems*". (Peralta et al., 2004) highlighted that: "*The concept of data freshness introduces the idea of how old is the data: Is it fresh enough with respect to the user expectations? Has a given data source the more recent data?*".

Similarly, our goal is to measure the freshness of a term (interest) by revising the notion of term frequency and by adjusting it with a temporal-biased function. In fact, a user interest toward a specific topic declines as time goes on and new interests appear. Thus, we assume the closer the term is the current date S_C, the more its temporal frequency would be significant.

We use the temporal feature based on the Kernel Gaussian function as a temporal-biased function after proving its effectiveness by prior work in the field of term positioning (Lv and Zhai, 2009; Gerani et al., 2010).

$$K(S_C, S_j) = \frac{1}{\sqrt{2.\Pi}.\sigma}.\exp\left[\frac{-(S_C - S_j)^2}{2.\sigma^2}\right] \tag{5.4}$$

where σ is the interpolation coefficient, S_C is the current date and S_j is a prior date.

In this approach, we believe that time factor is an important factor that determines an interest relevance. In fact, time-sensitive profile is able to detect changes in the user's behaviour and thus enhance the latest user preferences that are used just before personalization.

Fig. 5.1 illustrates three terms distributions using first a simple cumulative term frequency of three terms (see Fig. 5.1-a), compared with their revised cumulative frequency using Kernel (Fig. 5.1-b).
We notice that term TF1 starting with high frequency (Fig1-a) its kernel

version (CF1) increases slowly. However, term TF3 starting from low frequency (0 in this case), continue to increase until it reaches the same cumulative frequency than TF1. Its kernel version (CF3) overpasses CF1. Term TF2 which has a uniform distribution continue to increase uniformly.

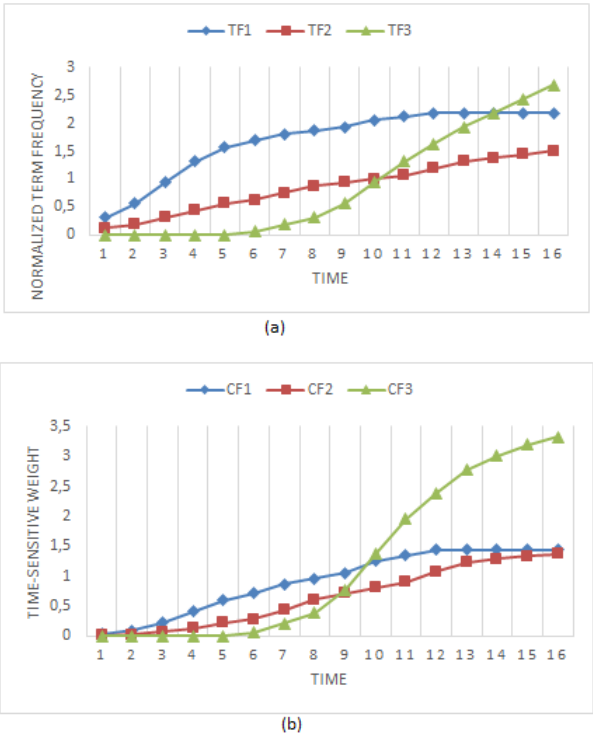

FIGURE 5.1: Example of terms distribution using cumulative term frequency (a) and a kernel version of the frequency (b)

Algorithm 1 summarized the profiling strategy proposed in our thesis:

Data: Documents: set of users' content, Terms: set of tweet' terms, S_j: activities' days
Result: $\overrightarrow{U}$: User profile vector

set $S_C \leftarrow$ Current Date ;
for *all d in Documents* **do**
  for *all t in Terms* **do**
    for *each day* S_j **do**
      W(t) $\leftarrow nTF(t_k)^{S_j}.K(S_C, S_j)$
    end
  end
end
$\overrightarrow{U} \leftarrow$ (t,W(t)) ;

Algorithm 1: User Profile Construction

5.3.2 Reranking

A retrieval model describes a computational process allowing ranking documents according to their relevance. In our work, relevant documents should satisfy not only the *topical relevance* but also be adapted to the user's interests.

The time-sensitive based user profile can be further refined by smoothing it with the Webpage-Query similarity obtained to personalize search results for all user's queries during the search session.

In fact, Webpage-Profile and Webpage-Query are both measured using the cosine similarity. It consists of measuring the cosine angle between the Web page $\overrightarrow{WP} = (t_{wp1}, t_{wp2}, ..., t_{wpk})$ and the profile in one hand, and between the Web page and the query (topical similarity) on the other hand.

Finally, the search results are re-ranked as follows:

$$Score(\overrightarrow{U}, \overrightarrow{Q}) = \alpha.Sim(\overrightarrow{WP}, \overrightarrow{Q}) + (1-\alpha).Sim(\overrightarrow{U}, \overrightarrow{WP}) \quad (5.5)$$

where $Sim(\overrightarrow{WP}, Q)$ is the score obtained from the original results reflecting the matching between the query and the Webpage, and $Sim(\overrightarrow{U}, \overrightarrow{WP})$ denotes the user-Webpage similarity. We present in the following the personalization algorithm.

Data: $\overrightarrow{U}$: user profile vector ; $\overrightarrow{Q}$: query vector, $\overrightarrow{WP}$: Web-page vector
Result: User profile vector with coupled (term:weight)
Set α: smoothing parameter ;
for *each* $\overrightarrow{U_i}$ **do**
  for *each* $\overrightarrow{Q_k}$ **do**
    for *each* $\overrightarrow{WP_j}$ **do**
      CALCULATE *similarity*($\overrightarrow{U_i}$, $\overrightarrow{WP_j}$) $\leftarrow$ Cos($\overrightarrow{U_i}$, $\overrightarrow{WP_j}$) ;
      CALCULATE *similarity*($\overrightarrow{Q_k}$, $\overrightarrow{WP_j}$) $\leftarrow$ Cos($\overrightarrow{Q_k}$, $\overrightarrow{WP_j}$);
      Score $\leftarrow \alpha$. Cos($\overrightarrow{Q_k}$, $\overrightarrow{WP_j}$) + (1-α). Cos($\overrightarrow{U_i}$, $\overrightarrow{WP_j}$);
    end
  end
end
SORT $\overrightarrow{WP}$ by *Score* ;

Algorithm 2: Personalization algorithm

5.4 Experiments

In this Section, we investigate the impact of the time-sensitive user profile strategy in the context of personalized search. More specifically, we examine the impact of our proposed temporal pattern in improving the accuracy of the Web search. Accordingly, our aim is to analyze and compare our approach with state-of-the-art approaches.
The main goals of these experiments are:

- Study the user profile modeling based on time-sensitivity,
- Analyze how the proposed Time-Sensitive User Profile (TSUP), outlined in Section 5.3, affects personalization,
- Evaluate its performance in comparison to two non-time sensitive approaches and a time-sensitive one.

5.4.1 Data Set

In our experiments, we used Twitter[3], a famous social network and microblogging platform created and launched in 2006 by Jack Dorsey, Evan Williams, Biz Stone, and Noah Glass [4].

According to (Tom et al., 2015), users on twitter have moved from using it as keep up with the news and considering it as a real-time service, to use it as a way of staying in touch. People read, share and send tweets about not only their daily life but also their thoughts.
Table 5.1 gives details about this microblogging network. Those details show that there is a great potential on data provided in Twitter. We find that Twitter represents an interesting system that allow collecting data to create user profiles as users tweet on almost daily basis and spontaneously share content that interest them.

Feature	Value
Total number of registered Twitter users	695,750,000
Number of new Twitter users signing up everyday	135,000
Average number of tweets per day	58 million
Number of Twitter search engine queries every day	2.1 billion

TABLE 5.1: Twitter Statistics (Statistic-Brain, 2016)

We used the Twitter microblogging service to construct the user profile due to various motivations:

- **Activities Nature**
  Twitter is considered as a tool that allow people share their opinions and thoughts, *follow* other people and media accounts in order to be updated to things that interest them (e.g, technology, music, politics...)
- **Self Representation**
  Some users tweet about their everyday life and events they participated to. Sharing about their participation to events or about news is a spontaneous action reflecting users' interests.
- **APIs availability**
  Twitter provides rich APIs[5]. There are two types of API that can be easily used: RETS and Streaming. REST APIs allow reading data

[3] www.twitter.com
[4] https://en.wikipedia.org/wiki/Twitter
[5] http://dev.twitter.com

about users, their tweets and locations while Streaming APIs allow searching the public global stream generated in real time.

- **No Invasion of Privacy**
  Twitter accounts are publicly visible to all users unless users specify some extra privacy measures or make their profile protected.

Over a period of the first two weeks of December 2013, we crawled the microblogging system Twitter posts (tweets) to randomly select 800 users and extract their public 69000 tweets. The main details of our data set are presented in Table 5.2.

Number of Users	800
Period	01/12/2013 - 15/12/2013
Total Number of Tweets	69000
Average Number of Tweets per participant	86.25
Average Number of Tweets per participant per day	5.75

TABLE 5.2: Twitter Data Set Details

In order to collect this data set, we used *Twitter4J API* which is an unofficial Java library for the Twitter API[6]. It allows to easily integrate the Twitter API (Application Programming Interface) in any Java application. The library offers classes to manipulate the methods offered by the Twitter Streaming API that allow collecting tweets, retweets, followers and favorites of users.

5.4.2 Data Processing

Data processing is the first step consisting of the stop words removal, stemming and tokenization of documents and users' extracted terms thanks to *Apache Lucene* [7] tool.

Lucence is an open-source, high performance and scalable tool used in the information retrieval and extraction field. It is the main component of Apache Solr project and implemented in Java. It can be used in large databases by searching and indexing any textual data.

This processing is used for both profiles and documents as shown in Fig. 5.2.

[6] www.twitter4j.org
[7] www.lucene.apache.org

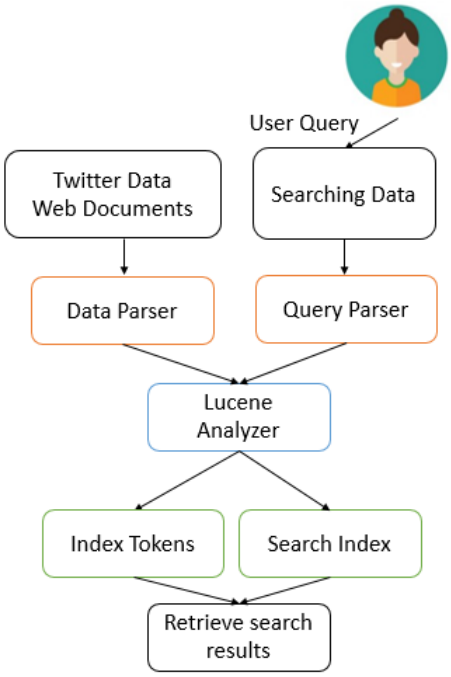

FIGURE 5.2: Data Processing using Lucene

We note that this tool offers a variety of *Analyzers* allowing to examine the text data and generate a token stream. We used the *Standard Analyzer* which is a bilangual analyzer for both English and French using a sophisticated grammar and a variety of filters *StandardTokenizer*, *StandardFilter*, *LowerCaseFiter*, *StopFilter*, and *PorterStemFilter* as a stemmer.

5.4.3 Evaluation Protocol

We select a unique query for each user profile related to his areas of interests defined on Twitter totaling 800 queries. Our queries are randomly selected from the online Twitter categories of interests (computer science, politics, chemistry, ...).

In order to submit each query, we used *Google Web Search API*. The API enables displaying results from Google searches including text and URL results. We select the top 100 documents per query and rerank as described in Section 5.3.2.

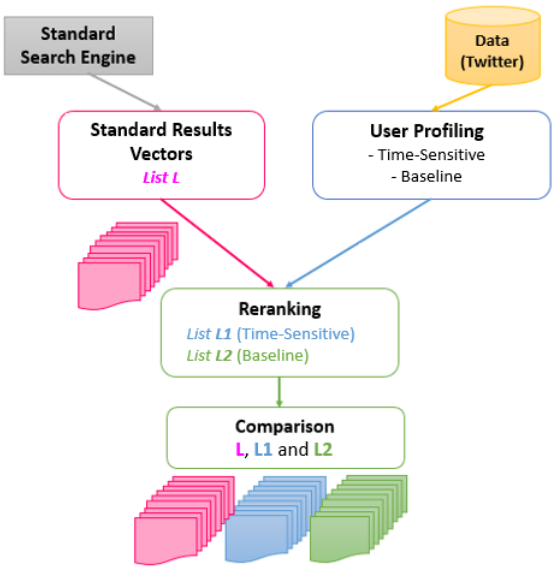

FIGURE 5.3: Overview of the Evaluation Methodology

In the Fig. 5.3, we present a general overview of our evaluation methodology. The evaluation methodology consists of three main steps: First, we collect data about the user in order to represent his profile as outlined before. Second, we submit the user's query to a standard search engine and select the top N relevant documents. Finally, we compare our approach with the standard results list, the one returned from weighting the user vector using only the frequency, and the one obtained after integrating the time.

5.4.4 Evaluation Measures

In order to measure the quality of the results, we use the *Normalized Discounted Cumulative Gain* NDCG@10, and Precision@10 for all the judged queries (Section 2.5.2). We considered any positive judgment as relevant.

Results were judged by 40 voluntary assessors with three levels of relevance, namely highly relevant (value equal to 2), relevant (value equal to 1) or irrelevant (value equal to 0). The assessors are graduate students in different fields, i.e., computer science, chemistry, tourism, electrical engineering, and medical. Each assessor evaluates the top-10 results of 20 entities (user, query, documents).

5.5 Results of the proposed experiments

In this Section, we lay out the findings of our analysis. First, we present results obtained by comparing our model with state-of-the-art approaches. Then, we try to specify the impact of growing information about the user's activities on the social Web and the enrichment with twitter-specific features.

5.5.1 Parameter Tuning

We used linear combination parameter α in order to adjust the importance of topical and temporal features of our model.

Fig. 5.4 presents the impact of α parameter for P@10. With $\alpha = 1$, there is no consideration of the temporal dynamics nor the user profile but only the topical relevance is taught into account. Inversely, $\alpha = 0$ corresponds to the topical based on the frequency.

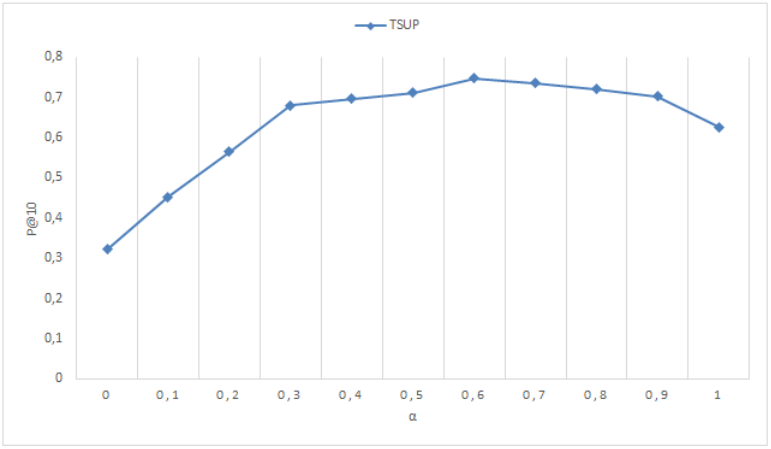

FIGURE 5.4: Parameter Tuning α

We note that at $\alpha = 0,6$ there is a peak of the P@10. We conclude that the topical relevance has a significant impact on the personalization performance that could be improved by integrating the user profile.

5.5.2 Baselines Comparison Results

We compare our time-based user profile with the non-personalized results and with the re-ranked list of documents returned by modeling the user profile. Thus, we used the following configurations:

- *TF.IDF*: Standard results obtained by Google search engine and weighted according to the TF.IDF scheme as presented in Section 5.3.2.
- *BM25*: Standard results weighted according to the BM25 scheme (Section 2.3.1).
- *nTF User Profile (nTF_{UP})*: user profile based only on the frequency of user activities.
- *Time-Sensitive User Profile (TS_{UP})*: our time-sensitive user profile based on merging the frequency and the freshness (Kacem et al., 2014).

From this comparison $(\alpha = 0.6, \sigma = 4)$, we obtained the values summarized in Table 5.3 where we use two metrics specified in Section 5.4.4.

	P@10	nDCG@10
Without Personalization		
TF.IDF	0,5787	0,4567
BM25	0,6022	0,5234
With Personalization		
nTF_{UP}	0,6268	0,5880
TS_{UP}	**0,7472**	**0,6256**

TABLE 5.3: Comparison results in terms of P@10 and nDCG@10

From the results presented in Table 5.3, we notice that our *TSUP* approach overcomes the results given by standard search engine as well as TF.IDF and BM25 schemes for both of nDCG@10 and P@10. The reranking based on time-sensitive user profile ensures an improvement of 16% to 23% for P@10 and an improvement of 25% to 42% for nDCG@10.

From our point of view, the reason of these values is the fact that the term frequency does not reflect the freshness of an interest but gives an overview of how often the user mentioned a term when interacting with the online social systems. However, standard search engines return relevant results to the user query's terms but they are indifferent to the users' interests especially when the queries are short (Jansen et al., 2000) or ambiguous (Cronen-Townsend and Croft, 2002).

Hence, the time-based user profile strategy defines current interests and needs of a user better than the non-time sensitive one. Furthermore, the standard search engine (e.g, Google) gives the same list of results without considering the user's individual needs because the ranking is based only on the matching of the document's terms to the query's keywords.

Consequently, merging both the freshness-feature and the term-frequency into our proposed weighting scheme has proved its effectiveness. The temporal dynamics allow considering the actual interests which are used to enhance the current search without overlooking the persistent interests and helps to personalize recurrent information needs.

5.5.3 Impact of User's Profile Information Amount

In order to better evaluate the influence of the temporal feature, we use the same personalization methodology to compare the time-sensitive user profile (TSUP) with the nTF-based user profile in terms of three profiles' temporal aspects namely:

- *Short-term profile*: all tweets extracted during the current day,
- *Long-term Profile*: all previous tweets except for those in the current session (before current day),
- *Single Profile*: all the recent and old tweets as a single user profile.

The reported results in Fig. 5.5 indicate that when we merge both of the interests into a single profile, we have a growing amount of profiling information that leads to better improvements in retrieval relevance. A single user profile that exploits all user's interests give better results than using profiles based solely on short- or long-term interests. Indeed, our approach outperforms the nTF approach with the three temporal aspects. These results reveal the usefulness of using all available information about the user and not only be limited to either recent or old activities. Users' interests are evolving over time but old interests could give an overview of the user's persistent preferences especially for users who are very active.

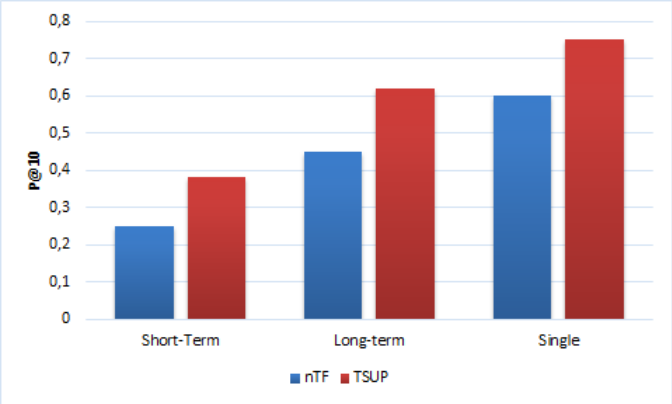

FIGURE 5.5: Comparison of mean of P@10

5.6 Discussion

Users are not always willing to express explicitly their interests in detail within forms. It is true that registration forms allow having data relevant to the user profile but are static and do not necessarily reflect the user's different interests, thoughts and preferences that are usually evolving.

In this chapter, we investigated how to implicitly construct the user profile extracted from his activities on the microblogging network Twitter. The profile is based on merging the long-term and short-term interests and used to improve the user experience with information retrieval systems when submitting a query.

We thus gave answers to the research questions raised at the beginning of this chapter as follows:

- The generic time-sensitive user profile (Section 5.3) can detect user preferences and be applied to personalized search results.It is based on considering both the content and temporal features.
- The user current and recurrent interests can be used both at different scales based on their freshness. We have studied the impact of using short-term and long-term profiles to optimally contribute to gains in relevance through search personalization and proved that they perform better in combination rather than using each one apart. Temporal dimension is an important dimension that allow understanding

the changes of user behavior and has a significant impact on improving the user search experience.

5.7 Conclusion

In this chapter, we explored the problem of personalized search and developed a user-modeling framework for Twitter microblogging system. In fact, the integration of the social data on the user model is accurate and efficient because people are likely to write a blog or bookmark a Webpage about something that interests them.

Furthermore, we investigate how the temporal-based user profile influences the accuracy of personalized search. We used a vector-based representation that takes into account the temporal-frequency measured by merging the term frequency and the freshness of each keyword using the Kernel function.

We find encouraging results when we compared our approach to two non-temporal sensitive approaches: the standard search engine Google (TF.IDF and BM25) and the user profiling using the Normalized Term Frequency scheme. In addition, we analyzed the aggregation of the current and recurrent interests. We found that increasing amount of profiling information yields to greater improvement in retrieval performance.

The evaluation using social media ensured significant results and could be extended to a different context: Searching History. Thus, in the following chapter, we present the research word conducte to evaluate the effectiveness of our approach using *Session Search*.

Chapter 6

Temporal Dynamics within Session Search

6.1 Introduction and Objectives

Usually a user interacts with a search engine by submitting a query describing his information need. A list of results are presented and he clicks on one or more results that interest him. When his information need has not been satisfied, the user reformulates the previous query with different manners (Jansen et al., 2000).

Recently, some works have moved from considering each submitted query independently (Liu, 2011) to taking into account the user behavior towards previous queries with the purpose of satisfying his current information need. A session search, as defined by (Boldi et al., 2008) is a sequence of queries issued by a single user within a specific time limit. Other definitions were proposed in the literature:

- *"A search session is all user activity within a fixed time window"* (Jones and Klinkner, 2008). The authors differentiate between search session, search goal and search mission and used a definition inspired

by *Merriam-Webster Dictionary* that defines a session as : *"A period of time that is used to do a particular activity"*.

- *"a series of interactions by the user towards addressing a single information need"*. (Jansen et al., 2007)
- *"an information retrieval task that involves a sequence of queries for a complex information need. It is characterized by rich user-system interactions and temporal dependency between queries and between consecutive user behaviors"*. (Luo et al., 2015)

Thus, an effective way to personalize a current query in a session resides in understanding the user's interests and preferences that can be modeled through a user profile and expressed in his previous interactions such as submitted queries, reformulated queries and clicked results.

Our aim is to improve sessions' results taking into consideration user's interactions under the assumption that recent performed ones are more related to the current needs than to the foregoing ones. In fact, a user reformulates a query in order to find new relevant information adapted to his current information need.

Specifically, we address the issue of leveraging user interactions with search engines in order to represent his profile as a vector of keywords where terms are weighted according to time-sensitivity. User's activities are presented in the form of a unique time-sensitive profile that merges both current and recurrent interactions giving more importance to recent ones.

First, we propose a temporal-frequency user profile that adjusts the term frequency according to its recency. Then, we chose *TREC Session* as an experimental framework in order to evaluate our proposed model. Session search provides test collections and evaluation measures that allow examining information retrieval over user sessions rather than current query only.

6.2 Time-Sensitive Session Search Model

We propose a user profile is represented as a vector of terms corresponding to the user interests and extracted from his browsing history. Our goal is explore user's past interactions and activities in order to enhance a current query's results.

In particular, the recent interests have more significant value than the old ones under the assumption that recent activities reflect better the user current needs. In fact, if the user submits different queries, we assume that the recent ones are more important because s/he didn't find relevant results from the previous ones. Precisely, we adopt a time-sensitive approach under the assumption that older frequent terms should not outperform current and not frequent ones.

6.2.1 Session Representation

Using session's interactions, we first collect keywords from past search queries and click-through documents. Then, we compute their weights by combining both their frequency and their appearing moment. We consider the browsing history but any other source does not affect our approach.

Formally, we consider a session search
$S = \{I_j = Q_j, \{D_j\}, \{C_j\}\}, Q_C \,|j = 1 : N\}$ where $\{I_j\}$ represents all previous interactions. Each interaction comprises a submitted query Q_j to which the search engine generates a list of documents $\{D_j\}$. A user clicks on a set of documents labeled $\{C_j\}$.

Each session contains a current query denoted Q_C for which we need to predict results $\{R\}$.

We define the user profile as a vector $\overrightarrow{U}$ of terms and their corresponding global weights W:

$$\overrightarrow{S} = (t_1 : W_1, t_2 : W_2, ..., t_n : W_n) \tag{6.1}$$

where $\{t_i \,|i = 1 : n\}$ are the terms forming the session profile, and W_i is the global weight described in the next session.

6.2.2 Content and Temporal Weighting

As a sessions contains issued queries and clicked documents, we assign a global weight W of a term t_i that is obtained by summing the weights obtained from previous queries $W_Q(t_i)$ and its clicked documents $W_C(t_i)$ as presented in equation 6.2. Clicked results and issued queries represent implicit feedback. Queries express users' information intent and clicked documents represent content that interest the user the most.

$$W(t) = \beta.W_Q(t) + (1 - \beta).W_C(t) \tag{6.2}$$

On the one hand, $W_Q(t)$ is computed through the following linear combination representing the weight obtained from all previous queries Q_i:

$$W_Q(t) = \sum_i nTF(t).K(Q_{Curr}, Q_i) \tag{6.3}$$

In equation 6.3, Q_{Curr} represents the current query, Q_i represents each previous query, nTF is the normalized-term-frequency of a term and $K(Q_{Curr}, Q_i)$ is its time-biased function that boosts term frequency of recent terms:

$$K(Q_{Curr}, Q_i) = \frac{1}{\sqrt{2.\Pi}.\sigma}.\exp\left[\frac{-({Q_{Curr}}^{ST} - {Q_i}^{ST})^2}{2.\sigma^2}\right] \tag{6.4}$$

We propose to use the Gaussian Kernel function which determines the weight of propagated terms between the current query and each previous one where

${Q_{Curr}}^{ST}$ represents the start time of a session's current query and ${Q_i}^{ST}$ represents the start time of each previous one submitted by the user during the same session. This accumulated frequency-biased weight can give the value of a term t at the current query Q_{Curr} by considering its positions at past queries Q_i favoring recent ones.

On the other hand, we measure the weight obtained from all clicked documents in a session $W_C(t)$ for each query Q_i:

$$W_C(t) = \sum_j TF.IDF(t).K({C_{Last}}^{ST}, {C_j}^{ST}) \tag{6.5}$$

where TF and IDF are the term-frequency and the inverse-document-frequency, $K({C_{Last}}^{ST}, {C_j}^{ST})$ represents its time-biased function as described in Equation 6.4 using the start-time of the j^th clicked document compared to the last one. If the document appears recently than it is more interesting for the user because s/he didn't find the information being sought in the previous documents and tends to explore new ones.

6.2.3 Linear combination

After measuring the resulting global weight of each term in the session-profile, we measure the score of each resulting document as follows:

$$Score(R) = \alpha.Sim(\overrightarrow{Q_{Curr}}, \overrightarrow{R}) + (1-\alpha).Sim(\overrightarrow{U}, \overrightarrow{R}) \tag{6.6}$$

where $Sim(\overrightarrow{U}, \overrightarrow{R})$ and $Sim(\overrightarrow{Q}, \overrightarrow{R})$ are the similarities between the user profile and the document result on the one hand, and between the result and the query on the other hand. α is the correlation variable. Both similarities are measured using the cosine function.

We give a sample session in Table 6.1. This session is composed of 7 queries, 4 of them has at least one clicked document that the user was interested in.

Previous query	Start-time	SAT Clicks
Q1. Scooter brands	79.932	clueweb12-1616wb-28-27881 clueweb12-0103wb-88-30226
Q2. Scooter brands reliable	229.262	clueweb12-0307wb-60-02121
Q3. Scooter	259.409	None
Q4. Scooter cheap	303.478	None
Q5. Scooter review	338.978	clueweb12-1616wb-28-27883
Q6. Scooter price	645.962	clueweb12-0002wb-43-35858
Q7. Scooter stores	690.053	None

TABLE 6.1: Session search example with current query "where to buy scooters"

Fig. 6.1 shows the distribution of queries' terms using cumulative frequency of words (a) as well as their distribution using cumulative time-based approach (b). The term "scooter"' has a uniform distribution for both approaches as it is used in all previous queries. In Fig. 6.1(a), terms "price",

"review" and "cheap" have the same value ($tf = 1$). However, in Fig. 6.1(b) terms appeared recently have higher value when using a temporal-based approach: $W(cheap) > W(brands) > W(reliable)$. In fact, we assume that a user changes keywords' queries when s/he has a need that has not been satisfied previously. Consequently, there is a need to consider the timing and to exploit a temporal distribution of added terms rather than considering only their occurrence.

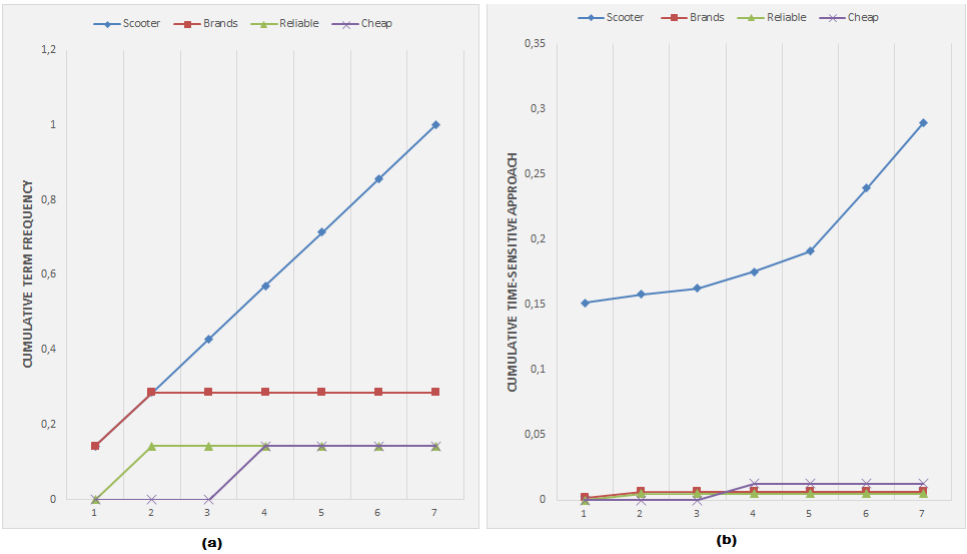

FIGURE 6.1: Distribution of queries' terms using term-frequency (a) and time-sensitive (b) approaches

We summarize in Algorithm 3 our personalization strategy within session search context.

Data: $S = t_1, t_2, ..., t_n$, n, $C_j \in C$: Clicked Documents, $Q_i \in Q$: Queries submitted in a session, $R_x \in R$: list of standard results returned in a session, Q_{curr}: Current query
Result: Personalized Ranking

Step 1: Measure the term's weight in a session profile
for *all* t_k *in* S **do**
  for *all* Q_i *in* Q **do**
    Measure the weight obtained from previous queries W(Q) (Equation 6.3) ;
    for *all* C_j *in* C **do**
      Measure the weight obtained from clicked documents W(C) (Equation 6.5) ;
    end
  end
  Measure the global weight combining W(t) combining W(C) and W(Q) (Equation 6.2) ;
end
$S \leftarrow$ (t,W(t)) ;
Step 2: Ranking personalization
for *all* R_x *in* R **do**
  Measure Sim_{Curr} the similarity between each result R_x and current queries Q_{Curr} ;
  Measure $Sim_{Session}$ the similarity between each result R_x and the session S ;
  Score (R) $\leftarrow$ Merge $Sim_{Session}$ and Sim_{Curr} ;
end
SORT R_x by *Score* ;

Algorithm 3: Time-Sensitivity in Session Search

6.3 Experimental Evaluation

In this Section, we investigate the impact of the time-sensitive user profile strategy in the context of session search using 2013 TREC Session track data. More specifically, we examine the impact of our proposed temporal pattern in improving the accuracy of the Web search. We particularly analyze how the proposed *Time-Sensitive Session Model*, described in Section 6.2, affects personalization and achieves better performances comparing to two baseline approaches namely the standard results returned for the current query without considering prior information, and the personalization approach using state-of-the-art approaches.

6.3.1 Dataset

To evaluate our work, we used 2013 TREC Session track. The track proposed 87 sessions used for evaluation. Each session has a topic describing

the aim of the search and covers historical queries and their issued times, ranked list of results, set of clicked URLs/ snippets and the time spent by the user visiting a URL with an average of 4.4 clicks.

More formally, each session consisted of the current query q_{curr} and the query session prior to the current query:

- the set of past queries in the session, q_1, q_2, ..., q_{curr-1}
- the ranked list of URLs for each past query,
- the set of clicked URLs/snippets and the time spent by the user reading the corresponding to each clicked url webpage.

Considering Fig. 6.2, in the left side, we find the number of sessions of a given length (in terms of total number of queries recorded). In the right side, we observe the amount of time spent in each session.

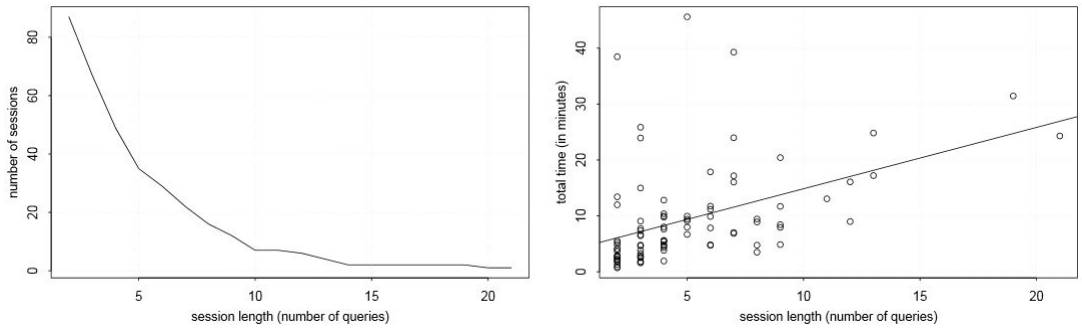

FIGURE 6.2: Session Length in 2013 TREC Session Track (Carterette et al., 2013)

Session track used the ClueWeb12 collection[1]. The full collection consists of roughly 730 million English language Web pages, comprising approximately 5TB of compressed data. We used Indri to perform retrieval within this collection, an indexing and retrieval component for the Lemur Toolkit[2] developed after collaboration between the universities of Massachusetts and Carnegie Mellon.
Session data are distributed in an XML (Extensible Markup Language) file format as follows:

```
<session num="1" starttime="0">
    <topic num="18">
        <desc>You are writing an article about face transplants. You want to
            know when the first full face transplant in the world was performed,
            in what country, city and hospital it was performed. Who was the
            lead doctor? what was the patient's name and age? And what caused
            the patient's facial problem?
        </desc>
    </topic>
    <interaction num="1" starttime="10.280644" >
        <query>wikipedia cosmetic laser treatment</query>
```

[1] http://lemurproject.org/clueweb12/
[2] http://www.lemurproject.org/

```
        <results>
            <result rank="1">
            <url>http://www.veindirectory.org/content/varicose_veins.asp</
                url>
            <title>Varicose Veins − Vein Treatment, Removal, Surgery
                Information</title>
          <snippet>... concern but can lead to more severe problems such as leg
              pain, leg swelling and leg cramps. View photos and find a varicose
              vein treatment center. ...</snippet>
        </result>
        <result rank="2">
          <url>http://www.peachcosmeticmedicine.com/treatments−Laser−
              and−IPL−hair−removal.html</url>
          <title>Laser and IPL hair removal − Treatments − Peach Cosmetic ...</
              title>
          <snippet>Laser hair removal served as Dr Mahony's introduction to
              cosmetic medicine back in 1999. ... Both our IPL and our laser offer
              skin chilling as part of the treatment. ...
            </snippet>
        </result>
        <result rank="10">
          <url>http://www.cosmeticsurgery10.com/index.html</url>
          <title>Cosmetic Surgery, Cosmetic Doctors, Cosmetic Physicians, and ...
              </title>
          <snippet>Cosmetic Surgery 10 is a resource that provides key
              information on cosmetic surgeries focusing on plastic surgeries,
              dermatology, cosmetic dentists and LASIK procedures.
            </snippet>
        </result>
      </results>
      <clicked>
        <click num="1" starttime="95.603468" endtime="120.565420">
            <rank>10</rank>
        </click>
      </clicked>
    </interaction>
    <currentquery starttime="252.659006">
      <query>uses for cosmetic laser treatment</query>
    </currentquery>
</session>
```

Each session is composed of queries regarding a specific topic. TREC Session Track 2013 contains 69 topics described in XML format as follows:

```
<topic num='1'>
    <desc>You are writing an article about the US civil war. Find relevant
        documents that talk about the following aspects: causes of the civil war,
        economic causes, battles in the civil war, consequences of the civil war,
        how greed affected the civil war, civil war effects today, what weapons
        were used during civil war, what rifles were used.</desc>
</topic>
```

6.3.2 Evaluation Protocol

We present in Fig. 6.3 the steps followed in order to compare our approach and evaluate the proposed model.

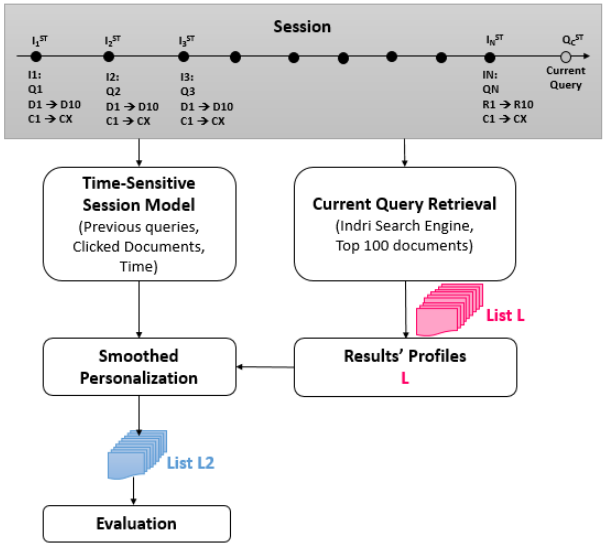

FIGURE 6.3: Temporal-based session personalization approach

We explain, thus, each step of those presented in Fig. 6.3.

- **Time-Sensitive Session Model**:
  We create a user profile for each session. We used the scoring approach described in details in Section 6.2 by choosing $\beta = 0,7$ giving the best value of precision after a series of test. We consider the start time of each query and each clicked result in order to give more value to recent submitted queries and seen results. We used Lucene[3] in order to index the session data, Lucene stop words removal and PorterStemFilter as stemmer.

- **Current Query Retrieval**:
  We submit the current query of each session using the Indri search engine and we selected top 100 results. Those results satisfy the content criterion and match the query's terms. As already mentioned, Indri is a search engine that provides state-of-the-art text search part of the Lemur Toolkit which is designed to facilitate research in information retrieval. We proceeded to the stop words removal and we used also Porter stemmer. Below, we present a current query retrieval sample of a current query "Where to buy scooters":

```
<parameters>
<memory>2560M</memory>
<index>IndexFile</index>
<query>
<number>1</number>
<text>where to buy scooters</text>
```

[3]https://lucene.apache.org/core/

```
</query>
<count>50</count>
<runID>runtest</runID>
<trecFormat>true</trecFormat>
<storeDocs>true</storeDocs>
<stemmer><name>porter</name></stemmer>
</parameters>
```

- **Results' Vectors Creation**:
  After getting Top-N documents' contents, we create their corresponding keywords-based vectors using *TF.IDF* as it is the most common scheme for Web documents' modeling. In order to insure novelty and diversity of results (Section 4.3.2), we have eliminated duplicate URLs and documents that have been clicked by the user.

- **Smoothed Personalization**:
  The similarity is measured using the cosine function between the document and query, on the one hand, in order to get the content score of a document. On the other hand, it is measured between the document and the profile allowing to get the temporal score of the document. Those similarities are aggregated linearly as described in Equation 6.6 by setting, after a set of experiments, $\alpha = 0,6$, $\beta = 0,7$ and $\sigma = 4$.

- **Relevance Evaluation**:
  We used judgment values provided by Session Track:

  - -2 for spam document (i.e. the page does not appear to be useful for any reasonable purpose; it may be spam or junk.);
  - 0 for not relevant (i.e. the content of this page does not provide useful information on the topic, but may provide useful information on other topics, including other interpretations of the same query);
  - 1 for relevant (i.e. the content of this page provides some information on the topic, which may be minimal; the relevant information must be on that page, not just promising-looking anchor text pointing to a possibly useful page);
  - 2 for highly relevant (i.e. the content of this page provides substantial information on the topic);
  - 3 for key, (i.e. the page or site is dedicated to the topic; authoritative and comprehensive, worthy of being a top result in a web search engine; typically, key pages are more comprehensive, have higher quality, and are from more trustworthy sources than the merely highly relevant page); and
  - 4 for navigational (i.e. this page represents a home page of an entity directly named by the query; the user may be searching for this specific page or site; there is often at most one page that deserves a Navigational judgment for an aspect).

6.3.3 Compared Models

In order to evaluate our retrieval system, we conduct experiments using several models:

- *None*: Standard results that are obtained by submitting only the current query without considering prior information (without personalization).
- *TF.IDF*: User profile approach based on Term-Frequency Inverse-Document-Frequency without considering the temporal factor.
- *BM25*: User profile approach based on BM25 (Robertson et al., 1995) scheme without considering the temporal factor.
- *TF.IDF+Ker*: Our time-sensitive user profile based on Kernel function and TF.IDF based frequency.
- *BM25+Ker*: Our time-sensitive user profile using BM25 to weight the content.

We propose to extend the experimental evaluation by adding two time-sensitive approaches using the following time-decay function:

$$Exponential_{Score} = \exp^{-\Delta T} \tag{6.7}$$

where ΔT represents *(1)* the difference between the start-time of the current query and the start-time of each previous query *(2)* the difference between the start-time of the last clicked document and each previously clicked one. . Thus, we consider the two following configurations:

- *TF.IDF+Exp*: time-sensitive user profile using the *Exponential Score* (Equation 6.7) with TF.IDF based frequency.
- *BM25+Exp*: time-sensitive user profile using the *Exponential Score* (Equation 6.7) with BM25 based frequency.

6.3.4 Measures

Based on the qrels provided by NIST, we evaluated the submitted configurations for the 87 queries used to evaluation.
First, we analyzed the *Mean Average Precision* in order to measure the average of relevant documents for each session, the *Precision* and *Recall* that are evaluated at a given cut-off rank, considering only the top *K* results returned by the system (k=10, 20). Given the *Precision* and *Recall*, we compute the F-Measure:

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6.8}$$

We used also the measures proposed by TREC Session track organizers:
The *Normalized Discounted cumulative gain (nDCG)* (Järvelin and Kekäläinen, 2002) using the the accumulated gain with the gain of each result discounted at lower ranks (nDCG@10 and nDCG@20).

The *Expected Reciprocal Rank (ERR)* is based on the cascade model of search. The cascade model assumes a user scans through ranked search results in order to evaluate whether the document satisfies the query of each session, and if it does, stops the search (Chapelle et al., 2009):

$$ERR = E(1/s) = \sum_{k=1}^{K} \frac{1}{k} p(q, d_k) \prod_{i=1}^{k-1} (1 - p(q, d_i)) \tag{6.9}$$

where s denotes the rank at which we stop, q is a query in a session, K is the number of returned documents, where the probability that document k satisfies the user query is given by the transform of the editorial grade assigned to the query-document pair : $p(q, d_k)$.

6.4 Results and Discussion

This Section provides the results and the impact of a dynamic representation of the user's search behavior.

6.4.1 Overall Performance Results

Table 6.2 highlights the results of Precision, Recall and F-Score at Rank 10 for all models.

	Precision@10	Recall@10	F-Score@10	% ↗
Without Personalization				
None	0,2010	0,1089	0,1503	42%
Without Time-Sensitivity				
TF.IDF	0,2556	0,1065	0,1664	39%
BM25	0,3500	0,1254	0,1846	25%
With Time-Sensitivity				
TF.IDF+Exp	0,3655	0,1505	0,2132	13%
BM25+Exp	0,3733	0,1574	0,2214	10%
TF.IDF+Ker	0,3971	0,1396	0,2066	16%
BM25+Ker	**0,4066**	**0,1759**	**0,2456**	-

TABLE 6.2: Performance comparison of our personalization approach using P@10, R@10 and F-Score@10 compared to different baselines. % ↗ indicates the improvement rate in terms of F-Score ($p < 0.05$ by a paired two-sided t-test)

From this Table, we notice that Standard Results (*None*) gives always the worst performance. These results are due to the fact that non-personalized approach consider only the current. It does not take into consideration any prior information such as previous queries an click-through-information.

In addition, *TSUP* achieves the best performance for F-Score@10 using *Kernal* or *Exponential*. These results confirmed our proposal that recent content contributes to improve personalized search. In fact, the aggregation of

users' past interactions within a session giving more importance to recent performed ones improves significantly the search precision.

The observations of Table 6.2 demonstrate, also, that both models based on *Kernal* temporal function improve personalized ranking. Specifically, the approach *BM25+Ker* gives the best results in terms of F-Score comparing to all models. It improves results from 42% to 25% compared to non-personalized approaches and from 10% to 16% compared to time-sensitive ones. Also, precision values are more important compared to Recall ones. This confirm that our approach detects user interests that are likely to be relevant rather than irrelevant.

We note also that considering the *BM25* weighting scheme to measure the thematic relevance gives better results comparing to *TF.IDF* scheme. In fact, considering the temporal function *Kernal*, *BM25+Ker* allows getting better performance than *TF.IDF+Ker*. This improvement is equal to 12%. Similarly, considering Exponential temporal function, *BM25* based scheme improves results by 4% comparing to TF.IDF scheme.

In Fig. 6.4, we detail the comparison between different models in terms of nDCG@10, nDCG@20 and MAP measures.
We observe that our proposed approach based on *TF.IDF* and *Kernel* function exceeds all models for both metrics nDCG@10 and MAP. This improvement is in order of 59% and 46% compared respectively to *TF.IDF* and *BM25*. As for time-sensitive approaches, the improvement varies between 41% and 45%.

Those results show clearly that our approach enhance results with significant relevance judgment and improve their ranking. For nDCG@20, our approach is very close to the model with best results which is *BM25+Exp* with 0,7610 compared to 0,7624.

In table 6.3, we compare our approach giving the best results **BM25+Ker** with the best results provided by TREC Session 2013 (Carterette et al., 2013). For both metrics, our approach is giving the best results which indicates the importance of using a temporal weight and not being limited to the thematic relevance.

Approach/Metric	nDCG@10	ERR@10
BM25+Ker	**0,4066**	**0,2525**
wdtiger2	**0.1952**	**0,1412**

TABLE 6.3: Comparison of our personalization approach compared to best run in TREC Session track 2013: wdtiger2[4]

[3]http://trec.nist.gov/pubs/trec22/appendices/session.html
[4]http://trec.nist.gov/pubs/trec22/appendices/session.html

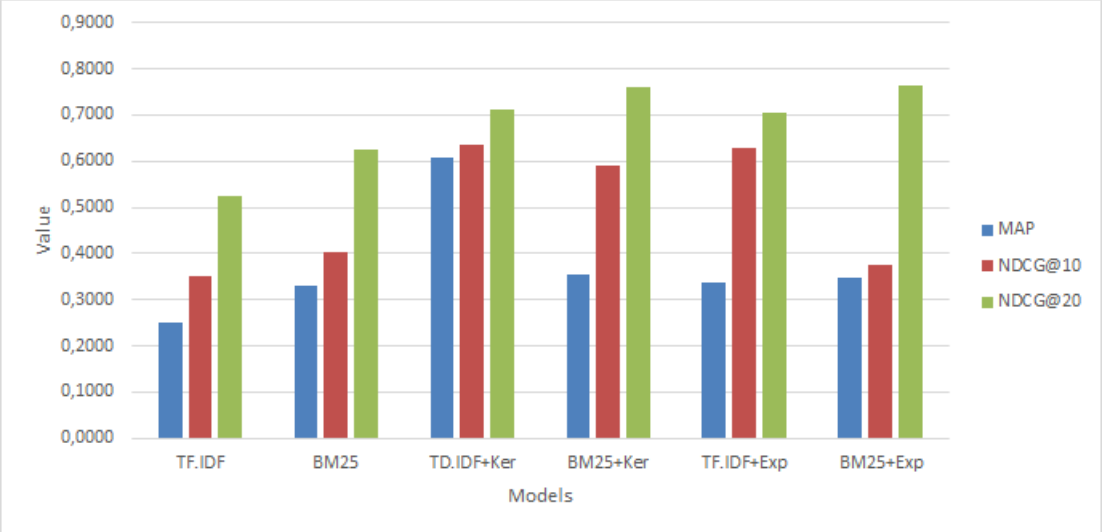

FIGURE 6.4: Comparison of compared models using nDCG@10, nDCG@20 and MAP

6.4.2 Features impact

We further evaluated in Fig. 6.5 the impact of each feature taken individually by considering past queries only *(Q)* ($\beta = 1$) and then clicked results only *(CL)* ($\beta = 0$). We take into account the Kernel Gaussian as temporal function because it gives the best results comparing to other models as discussed in Section 6.4.1.

For both configurations, we can see that the aggregation of those features indeed improves the search accuracy with an improvement of 48% using TF.IDF and of 50% using BM25 comparing to Q. Comparing to clicked documents-based model, the improvement in terms for *Precision@10* increases by 26% using TF.IDF and 37% using BM25.

In fact, previous queries contain few words comparing to clicked results. Terms in clicked documents are used to enrich queries' terms. A temporal distribution of those terms gives an overview of the moment of appearance in addition to how often they were used.

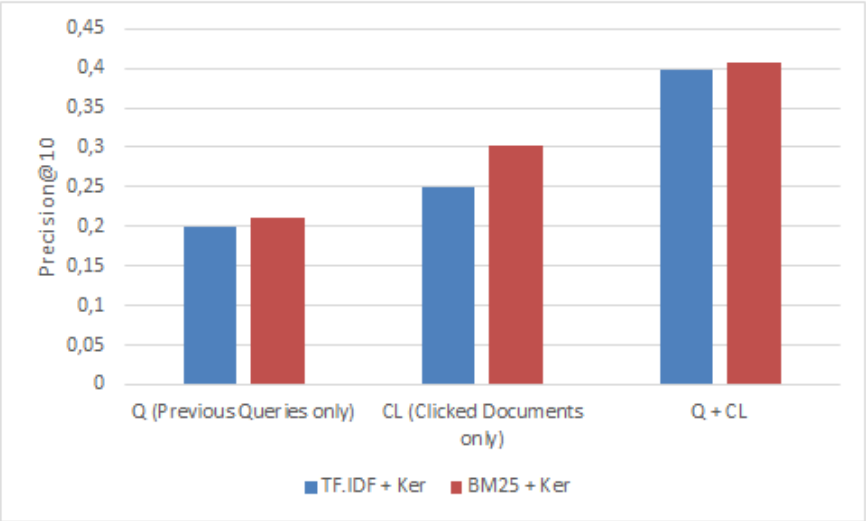

FIGURE 6.5: Impact of features on Precision@10 using the time-sensitive user profile using both Gaussian Kernel and Exponential

6.4.3 Dynamic Personalization impact

We now report the impact of the previous queries position on the performance of personalization. We found that the more queries we consider, the better the quality of the personalized rank. As the average of session length is 11.5, we consider the 11 submitted queries of each session and study their impact on MAP improvement.

From Fig. 6.6, we notice that the increase of MAP comes from recent history. These results enabled us to see how the recent interactions of a user affect the quality of the profile. In fact, terms used recently by the user reflects a new information need expressed in the most recent queries. This proves that the temporal feature has an impact on the improvement of the ranking.

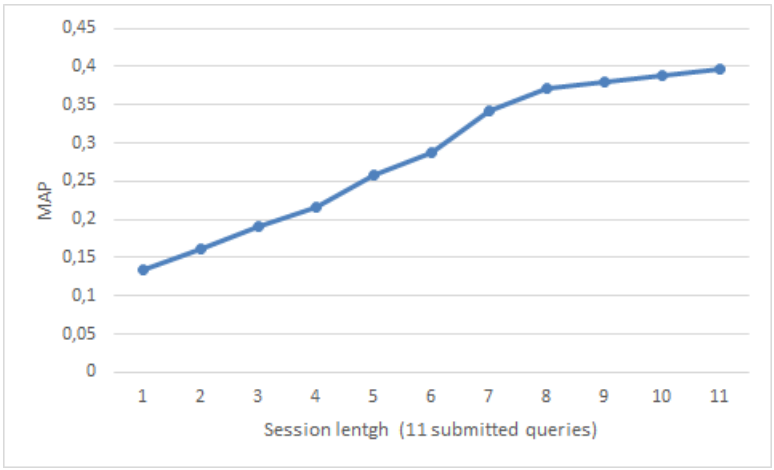

FIGURE 6.6: Query position impact on our personalization model

6.5 Conclusion

In this chapter, we investigated how the temporal-based user profile influences the accuracy of results in the context of session search. We proposed a time-sensitive approach that merges both frequency and freshness of user's actions thanks to the Kernel function. The vector-based representation takes into account the temporal -frequency of previous queries and clicked documents.

We compared our approach to non-temporal sensitive approaches: the standard results returned by Indri search engine, the user profile methods based on TF.IIDF and BM25 schemes, and a time-sensitive approach based on exponential function. Overall, we find promising results proving the impact of the temporal-frequency such as the query issue time and time spent visiting a Web-page.

In addition, we analyzed the aggregation of the current and recurrent information. We found that increasing amount of items appeared in recent queries yields to greater improvement in retrieval performance.

Chapter 7

Personalization for Enterprises

7.1 Introduction

Enterprises have moved from being product- or device-centred to be customer- or user-centred (Woods and Cortada, 2013). It is beneficial for companies, to consider the context of the user in order to improve search results or recommend items tailored to his interest and previous requests or visited items.

In this chapter, we present the relevance of the time-sensitive user profile in Web and mobile contexts within Orange Tunisia Corporation[1]. First, we present the corporation and its filed. Then, we discuss our contribution within the company and the effectiveness of our time-sensitive model. Finally, we highlight the advantages of this collaboration at different levels: the corporation, the Ph.D candidate and the research laboratory.

7.2 Orange Tunisia Corporation

Orange Tunisia is a telecommunications operator that acquired a license in Tunisia on 5th May 2010. It is considered as the leading 3G mobile operator with Divona Telecom, the second landline operator and the third mobile phone operator in Tunisia[2].

[1] www.orange.tn
[2] www.wikipedia.com - Orange Tunisia article

We have chosen Orange Tunisia corporation for this collaboration because since the year of the commercial launch, it has been committed to participate in the development of Tunisia through its businesses, technology, innovation, digital, by putting them at the service of the citizen. It is considered as an innovative, responsible and involved telecommunications operator.

7.3 Relevance of the research work for companies

From Orange Tunisia Corporation, this model ca be used at different levels, especially those in which we have contributed. First, we present the "Plus qu'un orpérateur" platform which is among the very few user-centered platform in Tunisia. Then, we present the relevance of our approach for a mobile application developed by Orange Developer Center (an innovative center for software development of Orange Tunisia Corporation).

7.3.1 Orange Tunisia, more than an operator

We have participated in the elaboration of the platform "*Orange Tunisie, plus qu'un opérateur*" through most of the steps (Nagaraj et al., 2010). It is consists in an innovative platform containing all the activities of the company related to the *innovation*, *corporate social responsibility* and *solidarity*.

FIGURE 7.1: *More than an operator* platform start page

We participated in the following steps:

- **Initiation Phase**
  During this step, we have participated in interviews with the different departments of Orange Tunisia because this project should meet the stakeholders expectations and needs. In addition, in this platform the stakeholders are from different domains : innovation, solidarity,

external communication, corporate social responsibility... A full description of the platform facets was provided, its feasibility, results, partners, and boundaries...

- **Definition Phase**
  This phase consists in defining the project functional and operational requirements. All the parties involved in the project has participated to this phase. Also, planing, risks, and costs were specified.
- **Design Phase**
  Design choices were identified through the requirements pointed out in the definition phase and adapted to the platform scope. Thus, the definitive platform hierarchy and Webpages' design are chosen by the project supervisors.
- **Development phase**
  During this phase, all the material and tools are established and a follow-up was required to verify completed project deliverables. Potential suppliers are indicated.
- **Implementation phase**
  The platform is than implemented involving developers, designers and contractors. Our work is limited to follow the contractor work and provide content inserted in the platform. Gathering, organizing and inserting the content with its different formats (text, images, videos...) are our main contribution to this step.
- **Testing phase**
  Testing the platform functionalities and corrective actions are also done in this step as well as content design, emplacement and display on different mobile platforms and different Web navigators.
- **Closing phase**
  More repairs can be done during this phase. Server problems can be faced because the platform can work for a small number of users and than face some problems when it is online. After verifying that, the launch of the platform was scheduled.

The time-sensitive user profile can be used to detect the user changes in the platform. In fact, when registering in the platform, the user specify its areas of interest among the corporation's fields : innovation, corporate social responsibility and solidarity. However, a user can register in the platform with a technological background and specify the "innovation" as an interest, but after navigating and discovering the other fields, we can be more interested in solidarity actions performed by the corporation.

Thus, through detecting the user navigation after his consent of course, we can recommend or personalize search in order to widen the result's and include both interests expressed explicitly and detected implicitly. The advantage of time-sensitivity is to be adapted to the user's need and also to take into consideration their changes over time.

7.3.2 Personalization for mobile applications enhancement

Mobile applications has been increasingly used in different fields (Ho and Kwok, 2002): culture, commerce, games... In order to improve the user navigation on mobile applications, the improvement of the recommendations and the personalization of interface and items' results have been introduced among companies priorities. Companies are not only willing to create useful applications with relevant content, but also to consider the enhancement of users' experience and provide them with customized items.

Orange Tunisia Corporation proposed a variety of mobile applications such as:

- **Mina7**: a mobile application that integrates all kinds of scholarships and academic events for students and doctoral candidates,
- **Karhbetna**: a mobile application for ride sharing allowing economic and ecological way of transportation,
- **Tunisia Passion**: a mobile application for the cultural tourism that offers a variety of services.

We are interested in the latter application "Tunisia Passion" that we took as an application example of our approach.

7.3.3 Tunisia Passion Mobile Application

Tunisia Passion, is an application dedicated to cultural tourism. The application aims to acknowledge the Tunisian heritage and tourism destinations, support the initiatives and the creations and deliver the concentrate of a country that moves and renews itself.

The application contains a variety of Tunisia regions and for each region, the user can choose one of the following functions (See Figure X):

- Discover: it consists in describing the region and its distinctive features.
- Stay : the user get a list of hotels in which he can stay.
- Relax: it contains a list of spa and tradition Tunisian bath addresses.
- Tour Visit: this function allow the user get a tour described in text and in the card so that s/he can have an idea about the region heritage; museums and most important places to visit.
- Savour: it contains the addresses about authentic tradition food and sweets.
- Offer: it marks all the shops of souvenirs and tradition jewelry and helps the user find ideas of gifts.

FIGURE 7.2: *Tunisia Passion* mobile application

7.3.4 Recommendation/ Personalization Improvement

We conducted preliminary analysis of Tunisia Passion's users in order to study their way of using the application. In fact, users checked the applications' items such as monuments, hotels, relax centers, restaurants and shops. They can rate up to five stars each of the checked items.

Our approach can be used with the explicit user profile as showed in Figure 7.3. In fact, we can use the ratings given by the user to the application's items in order to model his explicit profile. We also detect the changes in the explicit profile in order to merge it with the implicit one that is based on social media account of the user.

Thus, we can recommend items (such as restaurants and shops) related to the user's current preferences and interests. For example, if the user was interested in the Tunisian food a while ago and is now interested in the Italian cuisine, we can recommend items by alternating between both of the choices based on both current and old tastes.

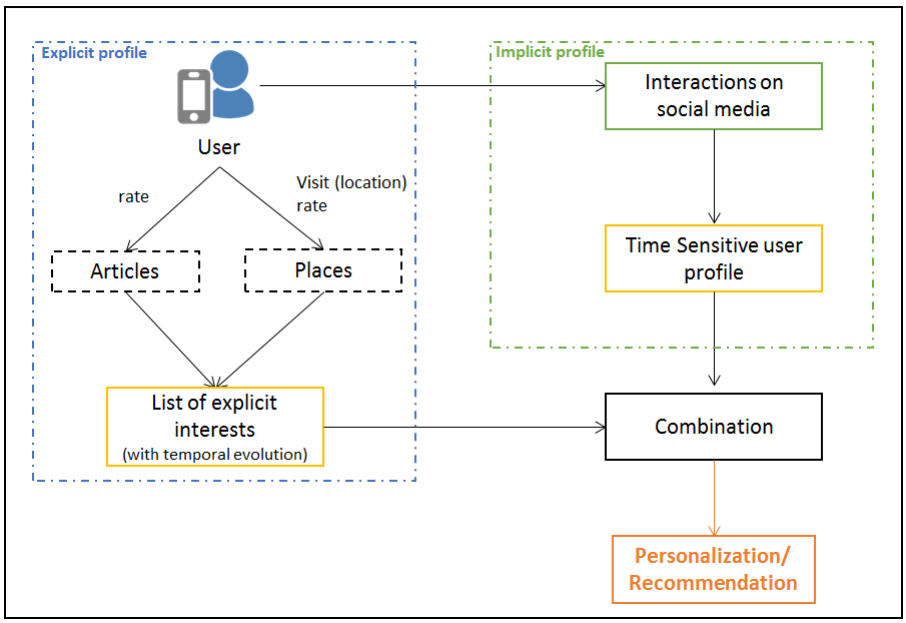

FIGURE 7.3: Recommendation/ Personalization based on time-sensitive user profile

7.4 Conclusion

In this chapter, we discussed the effectiveness of our research work in the Enterprise field. In fact, a research that is open to the enterprises could bring benefits to the three parts of the collaboration.

The company could get advantage of innovative research work that can bring relevant solutions for the products and services which can be enhanced in a way to improve the user experience and consider the user as the center of the company's offered services or products.
Customized items based on a time-sensitive user profile that are adapted to the single user prove the user importance and consideration instead of building personalization based only on the user history of actions and/or interactions.

It is valuable for each research laboratory to collaborate with companies through research projects and thesis. In fact, solutions addressed in the academic field can be directly applicable in the companies and thus research projects could be no longer limited to the theoretical aspect and could integrated concrete ones.

Finally, this collaboration is very constructive experience for the PhD candidate who faces different challenges and can benefit from both academic and professional environments.

Conclusion

Chapter 8

Conclusion

The integration of user profile in an information retrieval system is considered as a major challenge in setting up systems for adapting information to the individual user (personalization, recommendation, etc.). Such systems can be used in various applications: search engines optimization, commercial websites or Websites' interfaces personalization... It is modeled through various representations and can be extracted from various sources.

8.1 Summary of Contributions

In this dissertation, the user profile is implicitly constructed as a vector of weighted terms. More precisely, we are interested in the temporal distribution of a user's interests, preferences and behavior. Prior work discern the short-term and long-term profiles. The first one represents interests related to the user's current search activities. The second one represents user's persisting interests which are extracted from his entire search history. However, we find that the fact of discerning the short-term and long-term user profiles does not necessarily reflect the user's needs. For users who are not very active on social services, the short-term profile can eliminate relevant results which are more related to their personal interests. This is because their social activities are few and separated over time. In addition, for users who are very active, the aggregation of recent activities without ignoring the old interests would be very interesting because this kind of profile is usually changing over time. In fact, considering recent interactions in this context can eliminate relevant information about users' intended purpose of search.

The central hypothesis of this thesis is that time-sensitive user profile offers advantages over traditional profile weighting methods considering only the frequency of activities or only one temporal dimension (short-term or long-term). In order to prove this hypothesis, this thesis describes two user modeling strategies and experiments.

The main one, presented in Chapter 5, consists on Modeling a **Time- Sensitive User profile**. Precisely, we considered activities extracted from the microblogging system Twitter but any other source could be applicable considering each source' features.
We combined the content and temporal features by using a Kernel Gaussian function enhancing, thus, the recent content without ignoring the old one. It is considered as a revision to the classical notion of frequency in terms of mitigating the impact of high frequencies especially those appeared for a long time.

Chapter 6 describes the second model and the set of conducted experiments. It consists of considering the **Temporal Dynamics within Session Search** where a session is characterized by a current query, previous submitted queries and their corresponding results. We assume that recent submitted queries contain additional information explaining better the user intent and prove that the user hasn't found the information sought from previous submitted ones.

In Chapter 7, we described the advantage for companies and industry such as Orange Tunisia Corporation of innovative research work that can bring relevant solutions considering the user as the center of the company and adapting, thus, the offered services or products.

8.2 Findings

Conducting experiments based on a Twitter dataset and TREC Session track 2013 gave answers to research questions addressed in the beginning of this dissertations.

As for the first question *Which sources are considered sources of evidence for user profile modeling?*, we found that implicit extraction of user interests can identify users' preferences and interests. In fact, using both social media and searching history allow modeling a user profile that improved the accuracy of personalization.

Considering the second research question *How can current and recurrent interests be used in order to enhance personalization?*, experiments show that time-sensitive profile enhances the personalized search compared to non-personalization approaches using standard results returned by search engines (e.g. Google, Indri).

We extended the experiments and compared our model with other user modeling approaches namely *(a)* frequency-based user profile (TF.IDF and BM25 *(b)* time-sensitive user profile (Exponential-based temporal feature). In Chapter 6, we found that the use of a temporal feature based on either Kernel Gaussian or Exponential function improves search personalization using different measures such as Precision@k, nDCG@k and MAP... As for the content feature, the BM25 weighting scheme improves the topical relevance and enhances the performance of the user model.

Moreover, we analyzed the users' activities on social media in Chapter 5 and we identified that a growing amount of users' activities can bring additional information about their interests and preferences and thus improves the adapted results using Precision@10 and NDCG@10.

8.3 Future Work

In this thesis, we have considered the temporal user profile and study the impact of time-sensitive modeling in the context of personalized search.

- ***Utilization of information extracted from a user's network***
  While modeling the user profile using social media, we have employed Twitter microblogging system. We only used the textual data. As in other approaches that exploit linkage information (Younus, 2016; Mezghani et al., 2015), it could be useful to consider the user's connections.
  In fact, information extracted from his network could bring around semantically meaningful topics and extend his preferences as on social media platforms user's posts are influenced by followers or friends publishing activity. Also, public trends, events and news have been used in user profile modeling approaches (Gao et al., 2012; Abel et al., 2013; Bizid et al., 2015) as they can influence the user interests during a time slot.

- ***Integration of social signals***
  Furthermore, we can consider the social signals (Badache and Boughanem, 2015) such as *Like* on Facebook[1] or *+1 Mention* on Google+[2]. In fact, the addition of such features can enhance detecting interests especially that Facebook extended the "*Like*" button to add other reactions that can be used as feedback towards posts such as *"Angry"*, *"Love"*... This can be used for interests expansion and elimination of irrelevant interests that could be detected through textual data.

- ***Combination of explicit and implicit user profile***
  Data provided from users regarding their personal information and interests are considered precise and useful. In fact, the implicit inference of users' interests could be enhanced by explicit data directly elicited from them.

[1] www.facebook.com
[2] www.google.com

Bibliography

Abel, Fabian et al. (2011). "Analyzing Temporal Dynamics in Twitter Profiles for Personalized Recommendations in the Social Web". In: *Proceedings of the 3rd International Web Science Conference*. WebSci '11. Koblenz, Germany: ACM, pp. 1–8. ISBN: 978-1-4503-0855-7.

Abel, Fabian et al. (2013). "Cross-system user modeling and personalization on the Social Web". In: *User Modeling and User-Adapted Interaction* 23.2, pp. 169–209.

Abowd, Gregory D. et al. (1999). "Towards a Better Understanding of Context and Context-Awareness". In: *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing*. HUC '99. Karlsruhe, Germany: Springer-Verlag, pp. 304–307. ISBN: 3-540-66550-1.

Abrams, David, Ron Baecker, and Mark Chignell (1998). "Information Archiving with Bookmarks: Personal Web Space Construction and Organization". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '98. Los Angeles, California, USA: ACM Press/Addison-Wesley Publishing Co., pp. 41–48. ISBN: 0-201-30987-4.

Achemoukh, Farida and Rachid Ahmed-Ouamer (2014). "Representation and Evolution of User Profile in Information Retrieval Based on Bayesian Approach". In: *Foundations of Intelligent Systems: 21st International Symposium, ISMIS 2014, Roskilde, Denmark, June 25-27, 2014. Proceedings*. Ed. by Troels Andreasen et al. Cham: Springer International Publishing, pp. 486 –492.

Adomavicius, Gediminas and Alexander Tuzhilin (2001). "Expert-Driven Validation of Rule-Based User Models in Personalization Applications". In: *Data Mining and Knowledge Discovery* 5.1, pp. 33–58.

Allan, James et al. (2003). "Challenges in Information Retrieval and Language Modeling: Report of a Workshop Held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002". In: *SIGIR Forum* 37.1, pp. 31–47. ISSN: 0163-5840.

Allan, James et al. (2012). "Frontiers, Challenges, and Opportunities for Information Retrieval: Report from SWIRL 2012 the Second Strategic Workshop on Information Retrieval in Lorne". In: *SIGIR Forum* 46.1, pp. 2–32. ISSN: 0163-5840.

Alonso, Omar et al. (2011). "M.: Temporal Information Retrieval: Challenges and Opportunities". In: *In: 1st Temporal Web Analytics Workshop at WWW*, pp. 1–8.

Ankolekar, Anupriya et al. (2008). "The two cultures: Mashing up Web 2.0 and the Semantic Web". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 6.1. Semantic Web and Web 2.0, pp. 70 –75.

Badache, Ismail and Mohand Boughanem (2015). "Document Priors Based On Time-Sensitive Social Signals". In: *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March*

29 - April 2, 2015. Proceedings. Ed. by Allan Hanbury et al. Springer International Publishing.

Baeza-Yates, Ricardo (2009). "User Generated Content: How Good is It?" In: *Proceedings of the 3rd Workshop on Information Credibility on the Web*. WICOW '09. Madrid, Spain: ACM, pp. 1–2. ISBN: 978-1-60558-488-1.

Baeza-Yates, Ricardo A. and Berthier Ribeiro-Neto (1999). *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc. ISBN: 020139829X.

Bambia, Mariem and Rim Faiz (2015). "FRel: A Freshness Language Model for Optimizing Real-Time Web Search". In: *Intelligent Systems in Cybernetics and Automation Theory: Proceedings of the 4th Computer Science On-line Conference 2015 (CSOC2015), Vol 2: Intelligent Systems in Cybernetics and Automation Theory*. Ed. by Radek Silhavy et al. Cham: Springer International Publishing, pp. 207–216.

Bao, Shenghua et al. (2007). "Optimizing Web Search Using Social Annotations". In: *Proceedings of the 16th International Conference on World Wide Web*. WWW '07. Banff, Alberta, Canada: ACM, pp. 501–510. ISBN: 978-1-59593-654-7.

Barrett, Rob, Paul P. Maglio, and Daniel C. Kellem (1997). "How to Personalize the Web". In: *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. CHI '97. Atlanta, Georgia, USA: ACM, pp. 75–82. ISBN: 0-89791-802-9.

Baskaya, Feza, Heikki Keskustalo, and Kalervo Järvelin (2012). "Time Drives Interaction: Simulating Sessions in Diverse Searching Environments". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '12. Portland, Oregon, USA: ACM, pp. 105–114. ISBN: 978-1-4503-1472-5.

Beale, Russell (2006). "Mobile Blogging: Experiences of Technologically Inspired Design". In: *CHI '06 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '06. Montréal, Québec, Canada: ACM, pp. 225–230.

Becerra, Claudia Jeanneth, Sergio Gonzalo Jimenez, and Alexander F. Gelbukh (2013). "Towards User Profile-based Interfaces for Exploration of Large Collections of Items". In: *Decisions@RecSys*. Ed. by Li Chen et al. Vol. 1050, pp. 9–16.

Begg, Iain M., Joe Gnocato, and Wendy E. Moore (1993). "A Prototype Intelligent User Interface for Real-time Supervisory Control Systems". In: *Proceedings of the 1st International Conference on Intelligent User Interfaces*. IUI '93. Orlando, Florida, USA: ACM, pp. 211–214. ISBN: 0-89791-556-9.

Bennett, Paul N. et al. (2012). "Modeling the Impact of Short- and Long-term Behavior on Search Personalization". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '12. Portland, Oregon, USA: ACM, pp. 185–194. ISBN: 978-1-4503-1472-5.

Berry, Michael W., Susan T. Dumais, and Gavin W. O'Brien (1995). "Using Linear Algebra for Intelligent Information Retrieval". In: *SIAM Rev.* 37.4, pp. 573–595. ISSN: 0036-1445.

Biancalana, C. and A. Micarelli (2009). "Social Tagging in Query Expansion: A New Way for Personalized Web Search". In: *Computational Science and Engineering, 2009. CSE '09. International Conference on*. Vol. 4, pp. 1060–1065.

Billsus, Daniel and Michael J. Pazzani (2000). "User Modeling for Adaptive News Access". In: *User Modeling and User-Adapted Interaction* 10.2-3, pp. 147–180. ISSN: 0924-1868.

Bizid, Imen et al. (2015). "Identification of Microblogs Prominent Users During Events by Learning Temporal Sequences of Features". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. CIKM '15. Melbourne, Australia: ACM, pp. 1715–1718.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3, pp. 993–1022. ISSN: 1532-4435.

Boldi, Paolo et al. (2008). "The Query-flow Graph: Model and Applications". In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. Napa Valley, California, USA: ACM, pp. 609–618. ISBN: 978-1-59593-991-3.

Boughanem, M and J Savoy (2008). *Recherche d'information: état des lieux et perspectives*. Recherche d'informtion et web. Lavoisier, Paris: Hermes Science Publications.

Bouzeghoub, Mokrane (2004). "A Framework for Analysis of Data Freshness". In: *Proceedings of the 2004 International Workshop on Information Quality in Information Systems*. IQIS '04. Paris, France: ACM, pp. 59–67. ISBN: 1-58113-902-0.

Brin, S. and L. Page (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In: *Seventh International World-Wide Web Conference (WWW 1998)*.

Bruce, Bertram C. (1972). "A model for temporal references and its application in a question answering program". In: *Artificial Intelligence* 3, pp. 1 –25.

Buckley, C. et al. (1995). "New Retrieval Approaches Using SMART : TREC 4". In: *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pp. 25–48.

Buckley, Chris and Ellen M. Voorhees (2000). "Evaluating Evaluation Measure Stability". In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '00. Athens, Greece: ACM, pp. 33–40. ISBN: 1-58113-226-3.

Burke, Robin (2007). "The Adaptive Web". In: ed. by Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. Berlin, Heidelberg: Springer-Verlag. Chap. Hybrid Web Recommender Systems, pp. 377–408. ISBN: 978-3-540-72078-2.

Cai, Yi and Qing Li (2010). "Personalized Search by Tag-based User Profile and Resource Profile in Collaborative Tagging Systems". In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. CIKM '10. Toronto, ON, Canada: ACM, pp. 969–978. ISBN: 978-1-4503-0099-5.

Campos, Ricardo et al. (2014). "Survey of Temporal Information Retrieval and Related Applications". In: *ACM Comput. Surv.* 47.2, 15:1–15:41. ISSN: 0360-0300.

Carmel, David et al. (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. Hong Kong, China: ACM, pp. 1227–1236. ISBN: 978-1-60558-512-3.

Carterette, Ben et al. (2013). "Overview of the TREC 2013 Session Track". In: *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*.

Chapelle, Olivier et al. (2009). "Expected Reciprocal Rank for Graded Relevance". In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. CIKM '09. Hong Kong, China: ACM, pp. 621–630. ISBN: 978-1-60558-512-3.

Chen, Liren and Katia Sycara (1998). "WebMate: A Personal Agent for Browsing and Searching". In: *Proceedings of the Second International Conference on Autonomous Agents*. AGENTS '98. Minneapolis, Minnesota, USA: ACM, pp. 132–139. ISBN: 0-89791-983-1.

Chen, Zhenhong et al. (2013). "ICTNET at Session Track TREC 2013". In: *Proceedings of The Twenty-Second Text REtrieval Conference(TREC 2013)*. NIST Special Publication: SP 500-302.

Clarke, Charles L.A. et al. (2008). "Novelty and Diversity in Information Retrieval Evaluation". In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. Singapore, Singapore: ACM, pp. 659–666. ISBN: 978-1-60558-164-4.

Cleverdon, Cyril, Jack Mills, and Keen? Michael (1966). *Factors determining the performance of indexing systems volume 1. Design*. Tech. rep. Cranfield: College of Aeronautics.

Cronen-Townsend, Steve and W. Bruce Croft (2002). "Quantifying Query Ambiguity". In: *Proceedings of the Second International Conference on Human Language Technology Research*. HLT '02. San Diego, California: Morgan Kaufmann Publishers Inc., pp. 104–109.

Daoud, Mariam, Mohand Boughanem, and Lynda Tamine-Lechani (2009). "Detecting Session Boundaries to Personalize Search Using a Conceptual User Context". In: *Advances in Electrical Engineering and Computational Science*. Ed. by Sio-Iong Ao and Len Gelman. Dordrecht: Springer Netherlands, pp. 471–482.

Davis, Charles H. and Eva Kiewitt (1979). *Evaluating Information Retrieval Systems: The Probe Program*. Westport, CT: Greenwood Press.

Diaz, Fernando (2009). "Integration of News Content into Web Results". In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. WSDM '09. Barcelona, Spain: ACM, pp. 182–191.

Ding, Yi and Xue Li (2005). "Time Weight Collaborative Filtering". In: *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. CIKM '05. Bremen, Germany: ACM, pp. 485–492. ISBN: 1-59593-140-6.

Dou, Zhicheng, Ruihua Song, and Ji-Rong Wen (2007). "A Large-scale Evaluation and Analysis of Personalized Search Strategies". In: *Proceedings of the 16th International Conference on World Wide Web*. WWW '07. Banff, Alberta, Canada: ACM, pp. 581–590.

Dourish, Paul (2004). "What We Talk About when We Talk About Context". In: *Personal Ubiquitous Comput.* 8.1, pp. 19–30. ISSN: 1617-4909.

Dumais, Susan et al. (2003). "Stuff I've Seen: A System for Personal Information Retrieval and Re-use". In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. SIGIR '03. Toronto, Canada: ACM, pp. 72–79. ISBN: 1-58113-646-3.

Gao, Qi et al. (2012). "A Comparative Study of Users' Microblogging Behavior on Sina Weibo and Twitter". In: *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*. UMAP'12. Montreal, Canada, pp. 88–101. ISBN: 978-3-642-31453-7.

Garcia Esparza, Sandra, Michael P. O'Mahony, and Barry Smyth (2013). "CatStream: Categorising Tweets for User Profiling and Stream Filtering". In: *Proceedings of the 2013 International Conference on Intelligent User Interfaces*. IUI '13. Santa Monica, California, USA: ACM, pp. 25–36. ISBN: 978-1-4503-1965-2.

Gauch, Susan et al. (2007). "User Profiles for Personalized Information Access". In: *The Adaptive Web: Methods and Strategies of Web Personalization*. Ed. by Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 54–89.

Gerani, Shima, Mark James Carman, and Fabio Crestani (2010). "Proximity-based Opinion Retrieval". In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '10. Geneva, Switzerland: ACM, pp. 403–410. ISBN: 978-1-4503-0153-4.

Gils, B van and HA Proper (2003). "Profile-based retrieval on the World Wide Web". In:

Göker, Ayse and Hans I. Myrhaug (2002). "User Context and Personalisation". In: *6th European Conference ov Case Based Reasoning, ECCBR 2002, Aberdeen, Scotland, UK, September 4-7, 2002, Workshop Proceedings*, pp. 1–7.

Golder, Scott A. and Bernardo A. Huberman (2006). "Usage Patterns of Collaborative Tagging Systems". In: *J. Inf. Sci.* 32.2, pp. 198–208. ISSN: 0165-5515.

Gollapudi, Sreenivas and Aneesh Sharma (2009). "An Axiomatic Approach for Result Diversification". In: *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. Madrid, Spain: ACM, pp. 381–390. ISBN: 978-1-60558-487-4.

Guan, Dongyi, Sicong Zhang, and Hui Yang (2013). "Utilizing Query Change for Session Search". In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '13. Dublin, Ireland: ACM, pp. 453–462. ISBN: 978-1-4503-2034-4.

Guarino, Nicola, Claudio Masolo, and Guido Vetere (1999). "OntoSeek: Content-Based Access to the Web". In: *IEEE Intelligent Systems* 14.3, pp. 70–80. ISSN: 1541-1672.

Hannon, John, Mike Bennett, and Barry Smyth (2010). "Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches". In: *Proceedings of the Fourth ACM Conference on Recommender Systems*. RecSys'10. Barcelona, Spain: ACM, pp. 199–206. ISBN: 978-1-60558-906-0.

Haveliwala, Taher H. (2002). "Topic-sensitive PageRank". In: *Proceedings of the 11th International Conference on World Wide Web*. WWW '02. Honolulu, Hawaii, USA: ACM, pp. 517–526. ISBN: 1-58113-449-5.

Heppin, Karin Friberg (2012). *Test collections and the Cranfield Paradigm*.

Ho, Shuk Ying and Sai Ho Kwok (2002). "The attraction of personalized service for users in mobile commerce: an empirical study". In: *ACM SIGecom Exchanges* 3.4, pp. 10–18.

Hong, Liangjie, Aziz S. Doumith, and Brian D. Davison (2013). "Co-factorization Machines: Modeling User Interests and Predicting Individual Decisions in Twitter". In: *Proceedings of the Sixth ACM International Conference on Web*

Search and Data Mining. WSDM '13. Rome, Italy: ACM, pp. 557–566. ISBN: 978-1-4503-1869-3.

Inagaki, Yoshiyuki et al. (2010). "Session Based Click Features for Recency Ranking". In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI'10. Atlanta, Georgia: AAAI Press, pp. 1334–1339.

Ingwersen, Peter and Kalervo Järvelin (2005). *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 140203850X.

Jabeur, Lamjed Ben, Lynda Tamine, and Mohand Boughanem (2012). "Featured Tweet Search: Modeling Time and Social Influence for Microblog Retrieval". In: *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*. WI-IAT '12. Washington, DC, USA: IEEE Computer Society, pp. 166–173.

Jain, Mona et al. (2013). "Temporal Analysis of User Behavior and Topic Evolution on Twitter". In: *Big Data Analytics: Second International Conference, BDA 2013, Mysore, India, December 16-18, 2013, Proceedings*. Ed. by Vasudha Bhatnagar and Srinath Srinivasa. Cham: Springer International Publishing, pp. 22–36.

Jansen, Bernard J., Amanda Spink, and Tefko Saracevic (2000). "Real life, real users, and real needs: a study and analysis of user queries on the web". In: *Information Processing Management* 36.2, pp. 207 –227.

Jansen, Bernard J., Amanda Spink, and Vinish Kathuria (2007). "How to Define Searching Sessions on Web Search Engines". In: *Advances in Web Mining and Web Usage Analysis: 8th International Workshop on Knowledge Discovery on the Web, WebKDD 2006 Philadelphia, USA, August 20, 2006 Revised Papers*. Ed. by Olfa Nasraoui et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 92–109.

Järvelin, Kalervo and Jaana Kekäläinen (2002). "Cumulated Gain-based Evaluation of IR Techniques". In: *ACM Trans. Inf. Syst.* 20.4, pp. 422–446. ISSN: 1046-8188.

Jatowt, Adam, Yukiko Kawai, and Katsumi Tanaka (2011). "Calculating Content Recency Based on Timestamped and Non-timestamped Sources for Supporting Page Quality Estimation". In: *Proceedings of the 2011 ACM Symposium on Applied Computing*. SAC '11. TaiChung, Taiwan: ACM, pp. 1151–1158. ISBN: 978-1-4503-0113-8.

Jegou, Herve, Matthijs Douze, and Cordelia Schmid (2008). "Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search". In: *Proceedings of the 10th European Conference on Computer Vision: Part I*. ECCV '08. Marseille, France: Springer-Verlag, pp. 304–317. ISBN: 978-3-540-88681-5.

Jelinek, F and R Mercer (1980). "Interpolated estimation of markov source parameters from sparse data". In: *Workshop on Pattern Recognition in Practice*. North Holland, Amsterdam: ACM, pp. 381–397.

Jeon, Grace YoungJoo and Soo Young Rieh (2013). "The Value of Social Search: Seeking Collective Personal Experience in Social Q&A". In: *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*. ASIST '13. Montreal, Quebec, Canada: American Society for Information Science, 7:1–7:10. ISBN: 0-87715-545-3.

Jiang, Jiepu and Daqing He (2013). "Pitt at TREC 2013: Different Effects of Click-through and Past Queries on Whole-session Search Performance".

In: *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*.

Jiang, Jiepu, Daqing He, and Shuguang Han (2012). "On Duplicate Results in a Search Session". In: *Proceedings of The Twenty-First Text REtrieval Conference (TREC 2012)*.

Jiang, S. et al. (2015). "Author Topic Model-Based Collaborative Filtering for Personalized POI Recommendations". In: *IEEE Transactions on Multimedia* 17.6, pp. 907–918.

Jones, K.S. et al. (1975). *Report on the Need for and Provision of an Ideal Information Retrieval Test Collection*. British Library Research and Development reports. University Computer Laboratory.

Jones, Rosie and Kristina Lisa Klinkner (2008). "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs". In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. Napa Valley, California, USA: ACM, pp. 699–708.

Jurafsky, Daniel (2012). *Language Modeling*.

Kacem, Ameni, Mohand Boughanem, and Rim Faiz (2014). "Time-Sensitive User Profile for Optimizing Search Personlization". In: *User Modeling, Adaptation, and Personalization: 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings*. Ed. by Vania Dimitrova et al. Springer International Publishing, pp. 111–121.

Kacem, Ameni et al. (2016). "Towards Improving e-Government Services Using Social Media-Based Citizen's Profile Investigation". In: *Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance, ICEGOV 2016, Montevideo, Uruguay, March 1-3, 2016*, pp. 187–190.

Kacem, Ameni, Mohand Boughanem, and Rim Faiz (2017). "Emphasizing Temporal-based User Profile Modeling in the Context of Session Search". In: *The 32nd ACM Symposium on Applied Computing, SAC 2017, April 3-6, 2017, Marrakesh, Morocco.* ACM, pp. 925–930.

Kanhabua, Nattiya and Avishek Anand (2016). "Temporal Information Retrieval". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. Pisa, Italy: ACM, pp. 1235–1238. ISBN: 978-1-4503-4069-4.

Kanhabua, Nattiya, Roi Blanco, and Kjetil Nørvåg (2015). "Temporal Information Retrieval". In: *Found. Trends Inf. Retr.* 9.2, pp. 91–208. ISSN: 1554-0669.

Kaplan, Andreas M. and Michael Haenlein (2010). "Users of the world, unite! The challenges and opportunities of Social Media". In: *Business Horizons* 53.1, pp. 59 –68.

Keller, Richard M. et al. (1997). "Papers from the Sixth International World Wide Web Conference A bookmarking service for organizing and sharing URLs". In: *Computer Networks and ISDN Systems* 29.8, pp. 1103 –1114. ISSN: 0169-7552.

Kelly, Diane and Jaime Teevan (2003). "Implicit Feedback for Inferring User Preference: A Bibliography". In: *SIGIR Forum* 37.2, pp. 18–28.

Kiewitt, E.L (1979). *Evaluating information retrieval systems: the PROBE program*. Greenwood Press.

Kirsch, Sebastian Marius, Melanie Gnasa, and Armin B. Cremers (2006). "Beyond the Web: Retrieval in Social Information Spaces". In: *Advances in Information Retrieval: 28th European Conference on IR Research, ECIR 2006,*

London, UK, April 10-12, 2006. Proceedings. Ed. by Mounia Lalmas et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 84–95.

Konstan, Joseph A. et al. (1997). "Grouplens: Applying Collaborative Filtering to Usenet News". In: *COMMUNICATIONS OF THE ACM* 40.3, pp. 77–87.

Kotov, Alexander et al. (2011). "Modeling and Analysis of Cross-session Search Tasks". In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '11. Beijing, China: ACM, pp. 5–14. ISBN: 978-1-4503-0757-4.

Lawrence, Steve (2000). "Context in Web Search". In: *IEEE Data Engineering Bulletin* 23, pp. 25–32.

Li, Lei et al. (2014). "Modeling and broadening temporal user interest in personalized news recommendation". In: *Expert Systems with Applications* 41.7, pp. 3168 –3177.

Li, Lin et al. (2007). "Dynamic Adaptation Strategies for Long-Term and Short-Term User Profile to Personalize Search". In: *Advances in Data and Web Management: Joint 9th Asia-Pacific Web Conference, APWeb 2007, and 8th International Conference, on Web-Age Information Management, WAIM 2007, Huang Shan, China, June 16-18, 2007. Proceedings*. Ed. by Guozhu Dong et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 228–240.

Li, Runsheng et al. (2012). "Modeling user's temporal dynamic profile in micro-blogging using clustering method". In: *Information Science and Service Science and Data Mining (ISSDM), 2012 6th International Conference on New Trends in*, pp. 808–812.

Li, Xiaoyan and W. Bruce Croft (2005). "Novelty Detection Based on Sentence Level Patterns". In: *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. CIKM '05. Bremen, Germany: ACM, pp. 744–751. ISBN: 1-59593-140-6.

Lieberman, Henry (1997). "Autonomous Interface Agents". In: *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. CHI '97. Atlanta, Georgia, USA: ACM, pp. 67–74. ISBN: 0-89791-802-9.

Liu, Chang et al. (2010). "Analysis and Evaluation of Query Reformulations in Different Task Types". In: *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*. ASIS&T '10. Pittsburgh, Pennsylvania: American Society for Information Science, 17:1–17:10.

Liu, Tie-Yan (2011). *Learning to Rank for Information Retrieval*. Springer Berlin Heidelberg.

Lops, Pasquale, Marco de Gemmis, and Giovanni Semeraro (2011). "Content-based Recommender Systems: State of the Art and Trends". In: *Recommender Systems Handbook*. Ed. by Francesco Ricci et al. Boston, MA: Springer US, pp. 73–105.

Luo, Jiyun et al. (2015). "Designing States, Actions, and Rewards for Using POMDP in Session Search". In: *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*. Ed. by Allan Hanbury et al. Cham: Springer International Publishing, pp. 526–537.

Lv, Yuanhua and ChengXiang Zhai (2009). "Positional Language Models for Information Retrieval". In: *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '09. Boston, MA, USA: ACM, pp. 299–306. ISBN: 978-1-60558-483-6.

Ma, Zhongming, Gautam Pant, and Olivia R. Liu Sheng (2007). "Interest-based Personalized Search". In: *ACM Trans. Inf. Syst.* 25.1.

MacKay, David J.C. and Linda C. Bauman Peto (1994). "A Hierarchical Dirichlet Language Model". In: *Natural Language Engineering* 1, pp. 1–19.

Mahmood, Tariq and Francesco Ricci (2009). "Improving Recommender Systems with Adaptive Conversational Strategies". In: *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*. HT '09. Torino, Italy: ACM, pp. 73–82. ISBN: 978-1-60558-486-7.

Maina, Antony (2016). *20 Popular Social Media Sites Right Now*. Ed. by Small Business Trends. URL: http://smallbiztrends.com/2016/05/popular-social-media-sites.html.

Man, Ning, Chen Xunxun, and Wang Bo (2016). "Hierarchical user interest model based on large log data of mobile internet". In: *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*, pp. 1–5.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press. ISBN: 0521865719, 9780521865715.

Maron, M. E. and J. L. Kuhns (1960). "On Relevance, Probabilistic Indexing and Information Retrieval". In: *J. ACM* 7.3, pp. 216–244. ISSN: 0004-5411.

Matthias, Hagen et al. (2013). "Webis at TREC 2013-Session and Web Track". In: *Proceedings of The Twenty-Second Text REtrieval Conference (TREC 2013)*.

Matthijs, Nicolaas and Filip Radlinski (2011). "Personalizing Web Search Using Long Term Browsing History". In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11. Hong Kong, China: ACM, pp. 25–34. ISBN: 978-1-4503-0493-1.

Merriam-Webster Dictionary. Information Retrieval. Merriam-Webster.

Mezghani, Manel et al. (2015). "A Case Study on the Influence of the User Profile Enrichment on Buzz Propagation in Social Media: Experiments on Delicious". In: *New Trends in Databases and Information Systems - ADBIS 2015 Short Papers and Workshops, BigDap, DCSA, GID, MEBIS, OAIS, SW4CH, WISARD, Poitiers, France, September 8-11, 2015. Proceedings*, pp. 567–577.

Micarelli, Alessandro and Filippo Sciarrone (2004). "Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System". In: *User Modeling and User-Adapted Interaction* 14.2, pp. 159–200.

Micarelli, Alessandro et al. (2007). "Personalized Search on the World Wide Web". In: *The Adaptive Web: Methods and Strategies of Web Personalization*. Ed. by Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 195–230.

Michelson, Matthew and Sofus A. Macskassy (2010). "Discovering Users' Topics of Interest on Twitter: A First Look". In: *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*. AND '10. Toronto, ON, Canada: ACM, pp. 73–80. ISBN: 978-1-4503-0376-7.

Michlmayr, Elke and Steve Cayzer (2007). "Learning user profiles from tagging data and leveraging them for personal (ized) information access". In: *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference (WWW2007)*, pp. 1–7.

Mobasher, Bamshad (2007). "Data Mining for Web Personalization". In: *The Adaptive Web: Methods and Strategies of Web Personalization*. Ed. by Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 90–135.

Moens, Marie-Francine, Juanzi Li, and Tat-Seng Chua (2014). *Mining User Generated Content*. Chapman & Hall/CRC. ISBN: 1466557400, 9781466557406.

Morita, Masahiro and Yoichi Shinoda (1994). "Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval". In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '94. Dublin, Ireland: Springer-Verlag New York, Inc., pp. 272–281. ISBN: 0-387-19889-X.

Morris, Meredith Ringel, Jaime Teevan, and Katrina Panovich (2010). "What Do People Ask Their Social Networks, and Why?: A Survey Study of Status Message Q&a Behavior". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. Atlanta, Georgia, USA: ACM, pp. 1739–1748. ISBN: 978-1-60558-929-9.

Mostafa, Javed (2005). "Seeking better web searches". In: *Scientific American* 292.2, pp. 66–73.

Moukas, Alexandros and Pattie Maes (1998). "Amalthaea: An Evolving Multi-Agent Information Filtering and Discovery System for the WWW". In: *Autonomous Agents and Multi-Agent Systems* 1.1, pp. 59–88.

Moulahi, Bilel, Lynda Tamine, and Sadok Ben Yahia (2015). "When time meets information retrieval: Past proposals, current plans and future trends". In: *Journal of Information Science*.

MyYahoo! (1995). *My Yahoo Portal*. URL: www.my.yahoo.com (visited on 08/10/2016).

Nagaraj, Srinivasan, M Ramachandra, and J Ratna Kumar (2010). "Cyclic approach to Web based project management." In: *International Journal of Computers and Applications* 8.5, pp. 26–30.

Neubauer, Nicolas et al. (2007). "Distance Measures in Query Space: How Strongly to Use Feedback From Past Queries". In: *Proceedings of the IEEE/WIC /ACM International Conference on Web Intelligence*. WI '07. Washington, DC, USA: IEEE Computer Society, pp. 607–613. ISBN: 0-7695-3026-5.

Noll, Michael G. and Christoph Meinel (2007). "Web Search Personalization via Social Bookmarking and Tagging". In: *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*. ISWC'07/ASWC'07. Busan, Korea: Springer-Verlag, pp. 367–380. ISBN: 3-540-76297-3, 978-3-540-76297-3.

Orlandi, Fabrizio, John Breslin, and Alexandre Passant (2012). "Aggregated, Interoperable and Multi-domain User Profiles for the Social Web". In: *Proceedings of the 8th International Conference on Semantic Systems*. I- SEMANTICS '12. Graz, Austria: ACM, pp. 41–48. ISBN: 978-1-4503-1112-0.

O'Sullivan, Derry, Barry Smyth, and David Wilson (2003). "Explicit vs Implicit Profiling: A Case-study in Electronic Programme Guides". In: *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. IJCAI'03. Acapulco, Mexico: Morgan Kaufmann Publishers Inc., pp. 1351–1353.

Paliouras, Georgios (2012). "Discovery of Web User Communities and Their Role in Personalization". In: *User Modeling and User-Adapted Interaction* 22.1-2, pp. 151–175. ISSN: 0924-1868.

Park, Taemin Kim (1994). "Toward a Theory of User-based Relevance: A Call for a New Paradigm of Inquiry". In: *J. Am. Soc. Inf. Sci.* 45.3, pp. 135–141. ISSN: 0002-8231.

Pehcevski, Jovan and Benjamin Piwowarski (2009). "Evaluation Metrics for Structured Text Retrieval". In: *Encyclopedia of Database Systems*, pp. 1015–1024.

Peralta, Verónika, Raúl Ruggia, and Mokrane Bouzeghoub (2004). "Analyzing and Evaluating Data Freshness in Data Integration Systems". In: *Ingénierie des Systèmes d'Information* 9.5-6, pp. 145–162.

Philbin, James et al. (2008). "Lost in quantization: Improving particular object retrieval in large scale image databases". In: *In CVPR*.

Pitkow, James et al. (2002). "Personalized Search". In: *Commun. ACM* 45.9, pp. 50–55. ISSN: 0001-0782.

Ponte, Jay M. and W. Bruce Croft (1998). "A Language Modeling Approach to Information Retrieval". In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '98. Melbourne, Australia: ACM, pp. 275–281.

Porter, M. F. (1997). "Readings in Information Retrieval". In: ed. by Karen Sparck Jones and Peter Willett, pp. 313–316.

Pretschner, A. and S. Gauch (1999a). "Ontology based personalized search". In: *Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on*, pp. 391–398.

– (1999b). "Ontology based personalized search". In: *Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on*, pp. 391–398.

Qiu, Feng and Junghoo Cho (2006). "Automatic Identification of User Interest for Personalized Search". In: *Proceedings of the 15th International Conference on World Wide Web*. WWW '06. Edinburgh, Scotland: ACM, pp. 727–736. ISBN: 1-59593-323-9.

Ramasamy, Dinesh, Sriram Venkateswaran, and Upamanyu Madhow (2013). "Inferring User Interests from Tweet Times". In: *Proceedings of the First ACM Conference on Online Social Networks*. COSN '13. Boston, Massachusetts, USA: ACM, pp. 235–240. ISBN: 978-1-4503-2084-9.

Ricci, Francesco, Lior Rokach, and Bracha Shapira (2011). "Introduction to Recommender Systems Handbook". In: *Recommender Systems Handbook*. Ed. by Francesco Ricci et al. Boston, MA: Springer US, pp. 1–35.

RkjIslam (2016). *Top Microblogging sites list and its advantage*. Ed. by Mizmizi Blog. URL: http://mizmizi.com/top-microblogging-sites-list/ (visited on 10/03/2016).

Robertson, S. E. and S. Walker (2000). "Okapi/Keenbow at TREC–8". In: *The Eighth Text REtrieval Conference (TREC–8)*. Gaithersburg, MD: NIST, 151–162.

Robertson, S.E. et al. (1995). "Okapi at TREC-3". In: *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pp. 109–126.

Robertson, Stephen E. and Karen Sparck Jones (1976). "Journal of the Association for Information Science and Technology". In: pp. 129–146.

– (1988). "Document Retrieval Systems". In: ed. by Peter Willett. London, UK, UK: Taylor Graham Publishing. Chap. Relevance Weighting of Search Terms, pp. 143–160. ISBN: 0-947568-21-2.

Rocha, E et al. (1970). “The concept of "relevance" in information science : a historical review”. In: *Introduction to information science*. Introduction to information science, pp. 111–151.

Romero, Cristóbal et al. (2007). “Personalized Links Recommendation Based on Data Mining in Adaptive Educational Hypermedia Systems”. In: *Creating New Learning Experiences on a Global Scale: Second European Conference on Technology Enhanced Learning, EC-TEL 2007, Crete, Greece, September 17-20, 2007. Proceedings*. Ed. by Erik Duval, Ralf Klamma, and Martin Wolpers. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 292–306.

Rosen-Zvi, Michal et al. (2004). “The Author-topic Model for Authors and Documents”. In: *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. UAI ’04. Banff, Canada: AUAI Press, pp. 487–494. ISBN: 0-9749039-0-6.

Salton, G. (1971). *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Salton, G., A. Wong, and C. S. Yang (1975). “A Vector Space Model for Automatic Indexing”. In: *Commun. ACM* 18.11, pp. 613–620. ISSN: 0001-0782.

Salton, Gerard (1968). *A Comparison Between Manual and Automatic Indexing Methods*. Tech. rep. Ithaca, NY, USA.

Salton, Gerard and Michael J. McGill (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc. ISBN: 0070544840.

Sanderson, Mark (2010). “Test Collection Based Evaluation of Information Retrieval Systems”. In: *Foundations and Trends in Information Retrieval* 4.4, pp. 247–375.

Saracevic, T (1970). “The concept of "relevance" in information science : a historical review”. In: *Introduction to information science*. Introduction to information science, pp. 111–151.

– (1996). “Relevance reconsidered”. In: *Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)*. Copenhagen,Denmark, pp. 201–218.

Schiaffino, Silvia and Analía Amandi (2009). “Artificial Intelligence”. In: ed. by Max Bramer. Berlin, Heidelberg: Springer-Verlag. Chap. Intelligent User Profiling, pp. 193–216.

Shen, Xuehua and Cheng Xiang Zhai (2003). “Exploiting Query History for Document Ranking in Interactive Information Retrieval”. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. SIGIR ’03. Toronto, Canada: ACM, pp. 377–378. ISBN: 1-58113-646-3.

Shen, Xuehua, Bin Tan, and ChengXiang Zhai (2005). “Implicit User Modeling for Personalized Search”. In: *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. CIKM ’05. Bremen, Germany: ACM, pp. 824–831. ISBN: 1-59593-140-6.

Sieg, Ahu, Bamshad Mobasher, and Robin Burke (2004). “Inferring User’s Information Context from User Profiles and Concept Hierarchies”. In: *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004*. Ed. by David Banks et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 563–573.

Sivic, Josef and Andrew Zisserman (2008). “Efficient Visual Search of Videos Cast as Text Retrieval”. In: *IEEE Transactions on Pattern Analysis Machine Intelligence* 31, pp. 591–606.

Snášel, Václav et al. (2010). "User Profiles Modeling in Information Retrieval Systems". In: *Emergent Web Intelligence: Advanced Information Retrieval*. Ed. by Richard Chbeir et al. London: Springer London, pp. 169–198.

Sontag, David et al. (2012). "Probabilistic Models for Personalizing Web Search". In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. WSDM '12. Seattle, Washington, USA: ACM, pp. 433–442. ISBN: 978-1-4503-0747-5.

Speretta, M. and S. Gauch (2005). "Personalized search based on user search histories". In: *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pp. 622–628.

Statista (2016). *Statistics and Market Data on Social Media User-Generated Content*. Ed. by Statistia. URL: https://www.statista.com/markets/424/topic/540/social-media-user-generated-content/.

Statistic-Brain (2016). *Twitter Statistics*. Ed. by Statistic Brain. URL: http://www.statisticbrain.com/twitter-statistics/.

Sugiyama, Kazunari, Kenji Hatano, and Masatoshi Yoshikawa (2004). "Adaptive Web Search Based on User Profile Constructed Without Any Effort from Users". In: *Proceedings of the 13th International Conference on World Wide Web*. WWW '04. New York, NY, USA: ACM, pp. 675–684. ISBN: 1-58113-844-X.

Tamine-Lechani, Lynda, Mohand Boughanem, and Zemirli Nesrine (2008). "Personalized document ranking: Exploiting evidence from multiple user interests for profiling and retrieval". In: *Journal of Digital Information Management* 6.5, pp. 354–365.

Tan, Bin, Xuehua Shen, and ChengXiang Zhai (2006). "Mining Long-term Search History to Improve Search Accuracy". In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06. Philadelphia, PA, USA: ACM, pp. 718–723. ISBN: 1-59593-339-5.

Tang, X., Y. Xu, and S. Geva (2013). "Integrating Time Forgetting Mechanisms into Topic-Based User Interest Profiling". In: *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*. Vol. 3, pp. 1–4.

Tchuente, D. et al. (2010). "Visualizing the Evolution of Users' Profiles from Online Social Networks". In: *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, pp. 370–374.

Tchuente, Dieudonné et al. (2013). "A community-based algorithm for deriving users' profiles from egocentrics networks: experiment on Facebook and DBLP". In: *Social Netw. Analys. Mining* 3.3, pp. 667–683.

Teevan, Jaime, Susan T. Dumais, and Eric Horvitz (2005). "Personalizing Search via Automated Analysis of Interests and Activities". In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '05. Salvador, Brazil: ACM, pp. 449–456.

Tom, Rosenstiel et al. (2015). *How people use Twitter in general*. Ed. by Press Institute. URL: https://www.americanpressinstitute.org/publications/reports/survey-research/how-people-use-twitter-in-general/ (visited on 09/15/2016).

Trajkova, Joana and Susan Gauch (2004). "Improving Ontology-based User Profiles". In: *Coupling Approaches, Coupling Media and Coupling Languages*

for Information Retrieval. RIAO '04. Vaucluse, France: LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, pp. 380–390.

Ustinovskiy, Yury and Pavel Serdyukov (2013). "Personalization of web-search using short-term browsing context". In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. CIKM '13. San Francisco, California, USA: ACM, pp. 1979–1988. ISBN: 978-1-4503-2263-8.

Vallet, David, Iván Cantador, and Joemon M. Jose (2010). "Personalizing Web Search with Folksonomy-Based User and Document Profiles". In: *Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010.Proceedings*. Ed. by Cathal Gurrin et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 420–431.

Van Rijsbergen, C. J. (1986). "A non-classical logic for information retrieval". In: *The Computer Journal* 6.29, pp. 481–485.

Vargas, Saúl and Pablo Castells (2011). "Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems". In: *Proceedings of the Fifth ACM Conference on Recommender Systems*. RecSys '11. Chicago, Illinois, USA: ACM, pp. 109–116.

Vickery, Graham and Sacha Wunsch-Vincent (2007). *Participative Web And User-Created Content: Web 2.0 Wikis and Social Networking*. Paris, France, France: Organization for Economic Cooperation and Development (OECD). ISBN: 9264037462, 9789264037465.

Vu, Thanh et al. (2015). "Temporal Latent Topic User Profiles for Search Personalisation". In: *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*. Ed. by Allan Hanbury et al. Cham: Springer International Publishing, pp. 605–616.

White, Ryen W., Paul N. Bennett, and Susan T. Dumais (2010). "Predicting Short-term Interests Using Activity-based Search Context". In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. CIKM '10. Toronto, ON, Canada: ACM, pp. 1009–1018. ISBN: 978-1-4503-0099-5.

Widyantoro, Dwi H. et al. (1997). "Alipes: A swift messenger in cyberspace". In: *Proceedings of AAAI Spring Symposium on Intelligent Agents in Cyberspace*, pp. 62–67.

Woods, John A and James Cortada (2013). *The knowledge management yearbook 2000-2001*. Routledge.

Xiang, Biao et al. (2010). "Context-aware Ranking in Web Search". In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '10. Geneva, Switzerland: ACM, pp. 451–458. ISBN: 978-1-4503-0153-4.

Xu, Shengliang et al. (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. Singapore, Singapore: ACM, pp. 155–162. ISBN: 978-1-60558-164-4.

Xu, Zhiheng et al. (2011). "Discovering User Interest on Twitter with a Modified Author-Topic Model". In: *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*. WI-IAT '11. Washington, DC, USA: IEEE Computer Society, pp. 422–429. ISBN: 978-0-7695-4513-4.

Yan, Qiang, Lanli Yi, and Lianren Wu (2012). "Human dynamic model co-driven by interest and social identity in the MicroBlog community". In: *Physica A: Statistical Mechanics and its Applications* 391.4, pp. 1540–1545.

Yin, Hongzhi et al. (2014). "A Temporal Context-aware Model for User Behavior Modeling in Social Media Systems". In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. SIGMOD '14. Snowbird, Utah, USA: ACM, pp. 1543–1554. ISBN: 978-1-4503-2376-5.

Younus, Arjumand (2016). "Use of Microblog Behavior Data in a Language Modeling Framework to Enhance Web Search Personalization". In: *Information Retrieval Technology - 12th Asia Information Retrieval Societies Conference, AIRS 2016, Beijing, China, November 30 - December 2, 2016, Proceedings*, pp. 171–183.

Zhai, Chengxiang and John Lafferty (2004). "A Study of Smoothing Methods for Language Models Applied to Information Retrieval". In: *ACM Trans. Inf. Syst.* 22.2, pp. 179–214. ISSN: 1046-8188.

Zhang, Sicong and Hui Yang (2013). "Applying the Query Change Retrieval Model on Session Search-Georgetown at TREC 2013 Session Track". In: *Proceedings of The Twenty-Second Text REtrieval Conference (TREC 2013)*.

Zhao, Zhe et al. (2015). "Improving User Topic Interest Profiles by Behavior Factorization". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Florence, Italy: ACM, pp. 1406–1416. ISBN: 978-1-4503-3469-3.

Zimmermann, Andreas, Andreas Lorenz, and Reinhard Oppermann (2007). "An Operational Definition of Context". In: *Modeling and Using Context: 6th International and Interdisciplinary Conference, CONTEXT 2007, Roskilde, Denmark, August 20-24, 2007. Proceedings*. Ed. by Boicho Kokinov et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 558–571.