Journal of Theoretical and Applied Information Technology
30th April 2024. Vol.102. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EXPLORING SOCIAL INFORMATION RETRIEVAL: A CRITICAL ASSESSMENT AND COMPREHENSIVE REVIEW

HAMID KHALIFI[1], SARA RIAHI[2], AZIZ BOUJEDDAINE[3], YOUSSEF GHANOU[4], HICHAME KABIRI[5],

[1]ANISSE research Team, Faculty of Sciences, Mohammed V University in Rabat, Morocco

[2]National School of Architecture in Rabat, Morocco

[3,4,5] Laboratory IA, Moulay Ismail University of Meknes, Morocco

E-mail: [1]h.khalifi@um5r.ac.ma, [2]riahisaraphd@gmail.com, [3]azizboujeddaine@gmail.com,

[4]y.ghanou@umi.ac.ma, [5]hichamme@outlook.fr

ABSTRACT

Nowadays, social networks are used daily by thousands of people who utilize the internet worldwide. This activity, which has become one of the main ways that people use the Internet, provides several social functions, including content sharing among people who share interests and conversational exchanges. Scholars have known for almost two decades that social networks are valuable resources for comprehending how different facets of information retrieval (IR) have developed. Large volumes of important information are produced by social networks. The analysis of this valuable data is missed by standard information retrieval systems that treat documents basing on their content rather than their social surroundings. The solution to this problem is Social Information Retrieval (SIR). Its models take into account social media platforms and use social material as a secondary source of information within the IR system to progress the quality and relevance of search results. For this purpose, we propose a novel approach involving query expansion that integrates social information from documents into the similarity calculation between the query and the document. Our method involves refining the original query by adding additional details, thereby broadening the search scope to produce more satisfactory results.

Keywords: *IR, Sir, Social Networking Platforms, Netizens-Created Content, Social Surroundings, Social Behaviors.*

1. INTRODUCTION

The World Wide Web, which was invented in the last decade of the 20th century and initially consisted of a set of browsing units made up of Hyper Text Markup Language (HTML) documents interconnected via web links, has undergone a remarkable transformation and has embraced the principles of collaboration, where every user is both a creator and consumer of information.

The social web has become a key component of Web 2.0. It has completely transformed how internet users connect, communicate, and share information. Through different social networking platforms, Users can now create, modify, and publish enormous amounts of social data, which has resulted in the rise of Netizens-Created Content (NCC) in online social platforms. This social data is typically transient, and personal, and takes various forms, including views, clicks, comments, social relationships, likes, shares, +1s, and more. Furthermore, social content is distinguished by a set of characteristics such as audience, reliability, reputation, source, and so on.

Users' information that is generated from social media platforms creates a subdomain of IR: the Social Information Retrieval (SIR). Its processes uses additional source of data to improve search results [1]. Annotations, clicks, and other social interactions are examples of social content that influence how web resources are ranked [2, 3, 4, 5, 6, 7].

The ultimate purpose of using social data, especially interactions (likes, dislikes, +1s, etc.), is to harness the collective contributions of users,

often referred to as the 'wisdom of the crowd,' to enhance IR performance.

In [8], James Surowiecki described the participation of social community of users into data generation. They contribute by reviewing, commenting, and annotating web resources (Figure 1); which help new users find information that are visually appealing and socially engaging.

The tremendous volume of social information generated by users reinvigorate significant issues in IR, particularly regarding the pertinent calculation systems used in IR.

While traditional IR systems determines a resource's relevance using criteria taken from its content; SIR systems use heterogeneous and varied social data and incorporate it into the relevance calculation model. However, modeling such systems is still a big challenge for researchers.

Thus, SIR is an developing field at the intersection of information retrieval and social media analysis. In recent years, the explosive growth of Netizens-Created Content on platforms like web communities, forums and microblogs has led to an increased need for effective techniques to search, retrieve, and rank relevant information within this vast social landscape. Current approaches in SIR harness the power of social interactions, Netizens-Created Content, and contextual information to enhance traditional information retrieval methods. These approaches take into account factors such as user preferences, social relationships, and the temporal dynamics of social data, providing more personalized and context-aware search results.

One key aspect of SIR is the utilization of social signals, such as likes, shares, comments, and user profiles, to improve search results. Techniques like query expansion, personalized ranking, and context-aware retrieval are employed to deliver more accurate and contextually relevant information to users. Additionally, the evolving landscape of SIR explores the challenges of real-time and dynamic social data, as well as the ethical considerations related to user privacy and data security. As the volume of social data continues to grow, SIR remains a dynamic and evolving field, continuously adapting to the ever-changing nature of online social interactions.

The article is structured as takes after: an overview of the literature in this area is given in Parts 2 and 3. The Part 4 details the utility and requirements of leveraging social interactions to enhance IR. Part 5 presents our query reformulation method, and finally, we wrap up this effort in Part 6 with the results and directions for future research.

2. SOCIAL INFORMATION ON THE WEB

Social information encompasses any data generated by participants in social networks and all content whose social users have contributed to its development on the web.

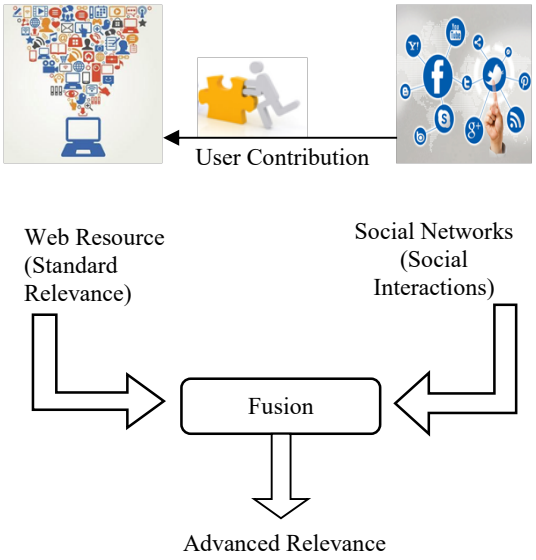

Figure 1: Leveraging Social Interactions for Enhanced Information Retrieval

In practice, a social platforms is a group of users connected to each other [9]. These users can be individuals, organizations, groups, and more. A social connection may involve two or more users engaged in activities like friendship, cooperation, alliances, or standard social information exchange. These internet users and their social connections evolve over time. New users and connections may form within the social platforms, while others may dissolve.

The following table outlines some of the most popular social media platforms:

Table 1: Popular Social Networks [10,11]

Designation	Creation Date	Active Users per Month	Description
Facebook	2004	2.8 milliards	"A social network allows you to publish information (statuses, photos, links, videos, etc.) and create pages and groups."
Twitter	2006	330 million	"A microblogging social network that allows users to send short messages,

			called 'tweets,' that do not exceed 140 characters."
GooglePlus	2011	500 million	"Google+ users could view updates from their contacts through circles via the Stream, which was similar to Facebook's news feed, but it was discontinued in 2019
LinkedIn	2002	774 million	"LinkedIn is a social network to be used in a business context. Users' profiles showcase their professional careers and enable them to specify their interests related to job opportunities, employment, hobbies, and other social aspects, all while sharing links, text, videos, and more."
Instagram	2010	1 billion	"Instagram allows users to share their photos and videos with their network of friends, provide likes, and leave comments on the images posted by other users."
TikTok	2016	1 billion	TikTok is a short video-sharing website where netizens can make and spread videos ranging from 15 to 60 seconds.
Snapchat	2011	280 million	Snapchat is a multimedia messaging app known for its self-destructing images and videos.

3. SOCIAL IR: CUTTING-EDGE RESEARCH

Traditional Information Retrieval (IR) has been altered to add numerous social characteristics in order to suit the information demands of netizens, this has resulted in the introduction of novel ways in information retrieval targeted at harnessing social data to improve the information retrieval process. And this is what we refer to as Social Information Retrieval (SIR).

Improving the traditional IR process involves utilizing social content as a secondary source of information, which can be achieved through three avenues [2]:

- Enhancing Indexing: This involves improving the way all term-document pairs are matched to measure the degree of their similarity.

- Query Refinement: Refining queries by incorporating supplementary data and information, which expands the user's query.

- Reordering Documents: Reordering search results depending on the user's social surroundings or other aspects of social interest.

3.1 Social Indexing

The insertion of tags to documents improves the retrieval of information [12, 13, 14, 15]

since they are regarded as helpful micro abstractions of documents [15]. On the other hand, papers containing few terms—for which typical indexing algorithms do not offer high-quality information retrieval—benefit greatly from the social information.

It has also been discovered that social data has been used in two unique ways to improve document representation [2]:

- The first approach involves enriching document content through dual indexing, incorporating both textuel substance and associated social substance, such as tags and all others social interactions. [13, 16, 14, 17, 7].

- The second approach involves creating personalized indexes to represent, comment on, and tag/label documents based on the personalized vocabulary of individual netizens [18, 19].

3.2 Query Refinement

Query refinement involves modifying the initial query by adding supplementary information to expand the search scope and obtain a more targeted and clearer query.

Several studies have been conducted in this regard, such as the approach introduced by Koolen et al. [20], which focuses on query expansion using Wikipedia as an external collection.

Additionally, the approach suggested by Li et al. [21] performs expansion processing on queries labeled as "weak" that do not yield enough relevant documents in searches. This approach has shown remarkable improvement, especially in terms of initially retrieved documents.

Furthermore, Bao et al. [22] propose the use of a social tag graph and web pages with arcs indicating the number of actors (see Figure 2). Through this schema, they suggest the iterative algorithm called Social-SimRank (SSR), which measures the degree of similarity for each pair of annotations.

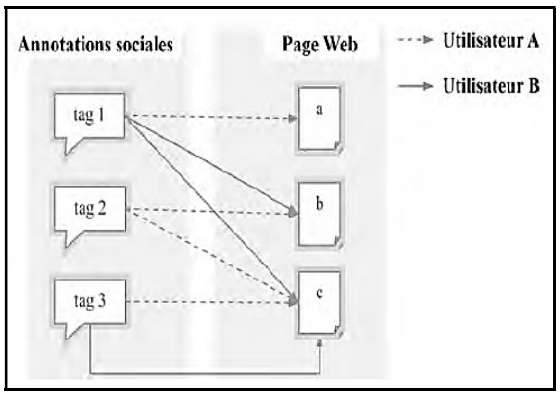

Figure 2: Illustration of Bao's SocialSimRank Graph [22]

3.3 Reordering of Results

Reordering results involves measuring the degree of similarity between document-query pairs to establish a new ranking. This ranking can be done in two categories of solutions that differ in how they utilize social information. The first category uses social information based on social relevance during ranking, while the second category utilizes social information to achieve personalized search results [2].

1) Ranking Based on Social Relevance:

The calculation of social relevance relies on social indicators that distinguish a document in terms of social interest, popularity, reputation, and more. Bao et al. [22] suggest the SocialPageRank algorithm, which measures the popularity of a given page based on the number of social tags to be used in ranking documents. Additionally, Yanbe et al. [23] introduced the SBRank function, which calculates the number of users who have annotated a given page and is used as a relevance calculation method in web searches.

2) Ranking Based on Social Surroundings:

The context, interests, and social habits vary from one user to another. Consequently, traditional IR models deliver documents ranked in the same way, which may not always align with the specific expectations and needs of the user.

To address this, various approaches have been proposed to rank documents differently according to the user's social surroundings. Some approaches use annotations provided by users to create their own profiles [24, 25]. Other works leverage favorited documents [26] as well as those that utilize social links [27].

Additional approaches have also been suggested, which leverage social surroundings and deliver search results with personalized rankings [28, 29, 30].

4. SOCIAL INTERACTIONS FOR IR IMPROVEMENT

In this section, we discuss and analyze existing approaches that use social interactions as supplementary sources to evaluate the relevance of resources. These approaches can be categorized into two groups: autonomous approaches that do not depend on time and those based on social interactions that depend on time.

4.1 Autonomous Approaches:

The research conducted by Cheng et al. [31] and Sally and David [32] is predicated on an analytical examination of YouTube video statistics and attributes that are associated with user activity during searches, including popularity, likes/dislikes, and the quantity of views a video has received. However, they did not use these social interaction parameters to improve the search system. Chelaru et al.'s research [7, 33] was centered on examining how social interactions affect YouTube search engine quality. By demonstrating a positive impact on improving the ranking of over 45% of searches, they illustrated the value of utilizing social annotations in addition to the similarity between the query and video titles. This was accomplished using "greedy feature selection algorithms" and ranking methods.

The goal of Karweg et al.'s work is to improve information retrieval by leveraging social surroundings [34]. They offered a method for calculating the level of social relevance by analyzing how much a user interacts with a web resource (clicks, reviews, comments, etc.) or using the credibility rate of each user measured through their social network graph, employing the PageRank algorithm to quantitatively calculate popularity.

A ranking model based on different social indices, as well as the level of interaction (for example, the number of plays for a specific soundtrack) on web resources, is proposed by Khodaei and Shahabi [35]. This model takes into account the connections between the person who started the search and the information creator. This model is based on extensive experiments conducted on a set of social data from an online radio,

significantly improving search result ranking compared to traditional approaches.

Buijs and Spruit [6] suggest a "Social Score Method" model based on a set of social interactions in social networks. The role of this social score is to select web resources that should be delivered primarily based on simple interaction calculations (shares, tweets, etc.). The results qualify this approach as a potential alternative to current systems such as the PageRank algorithm. In this approach, annotations are not considered social interactions because they require additional processing to be utilized. Instead, they use low-complexity social interactions, such as liking a post, which provides an easily measurable indicator of a document's performance.

Pantel et al. [36] presented the utility of social interactions (likes, dislikes, and shares) for improving information retrieval. They argue that a user's opinion on a web resource can satisfy the expectations of other users, and this can be achieved through various solutions, including social recommendations, personalized search, and the evaluation of web resource popularity.

Kazai and Milic-Frayling [37] propose a Social IR system that leverages various forms of opinions on books from a given collection (ratings, reviews, comments, etc.). Satisfactory results were obtained from experiments aimed at utilizing opinions in a traditional search system, using a collection containing more than 40,000 books and 250 queries with annotations and opinions from a group of users.

4.2 Time-Dependent Approaches:

The approaches presented in the previous section do not take into consideration the time factor related to interaction or when the web resource was created. There are not many achievements in information retrieval that focus on this aspect.

Inagaki et al. [38] suggest using the principles of "click-through rates" to measure the effectiveness of a web resource in terms of the ratio of clicks to impressions. This approach gives priority to documents that have a recent level of interest among users, resulting in a more efficient ranking of results compared to other models that rely on sorting based on what is recent.

Khodaei and Alonso [39] are not primarily interested in information retrieval. Instead, they aim to determine and measure user interests over time. They assume that the large volumes of data generated by users on social networks can be used to experimentally understand how information is produced and consumed by online actors over time. They categorize users' social interests into five categories: (recent, ongoing, seasonal, past, and random), and analyze users' activities on specific social platforms such as Twitter and Facebook. They also present solutions where this approach can be applied, including personalized information retrieval, information retrieval based on friendship relationships, and community-based information retrieval.

The approach by Ismail BADACHE et al. [2, 5] involves measuring the social relevance of a document using its associated social signals, either separately, where each signal denotes an importance index, or by combining these signals based on the type of relevance, with some signals related to popularity and others to reputation. To take these social indicators into account when calculating the importance rate, they rely on a language model that allows combining a document's interest and relevance after a given query.

5. USE OF SOCIAL SURROUNDINGS IN IR

Our suggestion is categorized under the query expansion axis, which is a method for changing the original query into a new one by adding more details. This broadens the search area and yields acceptable results.

In order to improve information retrieval, we seek to formulate a new type of query expansion by incorporating the social information of a given document into the similarity calculation between the query and the document (see Figure 3).

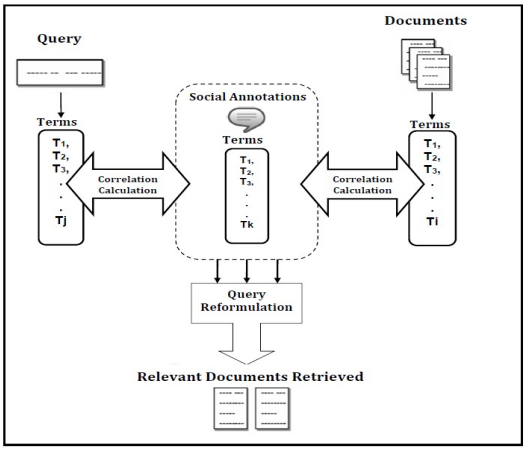

Figure 3: Proposed Social Information Retrieval Process

The information retrieval process with query reformulation Q occurs in the following steps:

- Step 1: Extraction of all terms from query Q

In the first step, the information retrieval process begins with the extraction of all terms from the user's query, denoted as Q. This initial query serves as the starting point for retrieving relevant documents from the information repository.

- Step 2: Identification of documents corresponding to each term in the query

Following the extraction of terms from the query, the system proceeds to identify documents that correspond to each term within the query. This step involves searching the information repository to find documents that contain the specific terms extracted from the user's query.

- Step 3: Extraction of annotations from the identified documents

Once the relevant documents are identified, the system extracts annotations from these documents. These annotations may include comments, tags, or other Netizens-Created Content associated with the documents. Annotations play a crucial role in providing additional context and relevance to the retrieved documents.

- Step 4: Correlation calculation using a language model Function

The system employs a "language model" function [40] to calculate correlations between terms. It calculates the correlation between each term in the document and each term in the annotations. Subsequently, it computes the correlation between each annotation term and each term in the original user query. These correlations help in understanding the semantic relationships between the terms.

- Step 5: Reformulate the initial query by including additional terms

Based on the calculated correlations, the system reformulates the initial user query by including additional terms. This query reformulation aims to broaden the scope of the search, making it more comprehensive and context-aware.

- Step 6: Execute the new query to obtain relevant documents

The newly reformulated query is then executed to retrieve documents that are now considered more relevant due to the expanded query. This step involves searching the information repository with the updated query to obtain documents that align more closely with the user's information needs.

- Step 7: Evaluate the performance of the approach compared to the PRF method

Finally, the performance of the approach is evaluated, often in comparison to the pseudo-relevance feedback (PRF) method. This evaluation helps assess the effectiveness of the query reformulation approach in improving the retrieval of relevant documents and enhancing the overall search experience.

6. CONCLUSION AND FUTURE WORK

The work presented in this article falls within the domain of information retrieval (IR), specifically in the context of social information retrieval. We addressed the challenge of effectively incorporating social interactions into the information retrieval process.

In this survey, we classified our work into three main aspects:

1. Defining social information on the web and social interactions.

2. Presenting the elements that require the inclusion of social content in the information retrieval (IR) process.

3. An overview of existing social IR approaches that exploit social interactions, categorized into two groups: those not time-dependent and those time-dependent.

The goal of exploiting social interactions is to leverage user opinions to significantly enhance IR performance and provide search results that align with users' needs and expectations.

However, upon analyzing the presented approaches, it is apparent that there are limited efforts that consider the time factor in information retrieval; and existing works that do consider this factor require further improvement, especially in terms of resource relevance calculation.

Considering textual content like comments, tweets, statuses, etc., remains challenging as they require additional processing for effective utilization.

It's worth noting that more than two-thirds of the referenced documents are post-2008, indicating the growing awareness among researchers regarding the importance of information contained in social

networks and the need for a new generation of IR models.

This work paves the way for several avenues of research in the field of social information retrieval. Our team is committed to refining our proposal to contribute to the enhancement of information retrieval systems in the future.

REFERENCES:

[1] Lamjed Ben Jabeur - "Leveraging social relevance: Using social networks to enhance literature access and microblog search" - Thèse de Doctorat - Université de Toulouse, Université Toulouse III-Paul Sabatier (2013)

[2] Ismail Badache - "Recherche d'Information Sociale : Exploitation des Signaux Sociaux pour Améliorer la Recherche d'Information" - Mémoire de thèse - Université Toulouse III-Paul Sabatier (2016)

[3] Thorsten Joachims - "Optimizing search engines using clickthrough data" In *ACM* (2002), 133–142

[4] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi, and WeiGuo Fan - "Optimizing web search using web click-through data" In *ACM* (2004), 118–126

[5] Ismail Badache and Mohand Boughanem - " Document priors based on timesensitive social signals" In *ECIR* (2015), Vol 9022

[6] Marco Buijs and Marco R. Spruit - "The social score - determining the relative importance of webpages based on online social signals" In *KDIR* (2014), 71-77

[7] S. V. Chelaru, C. Orellana-Rodriguez, et I. S. Altingovde - "Can social features help learning to rank youtube videos?" In *WISE* (2012), 552–566

[8] James Surowiecki - "The wisdom of crowds" In Anchor (2005)

[9] Robert A Hanneman and Mark Riddle - "Introduction to social network methods" - on-*line textbook* (2005)

[10] Thomas Coëffé - "Chiffres réseaux sociaux - 2023" (2023) available at : https://www.blogdumoderateur.com/chiffres-reseaux-sociaux/

[11] Wikipedia - "Instagram" available at : https://fr.wikipedia.org/wiki/Instagram

[12] Xiaoxun Zhang, Lichun Yang, Xian Wu, Honglei Guo, Zhili Guo, Shenghua Bao, Yong Yu, and Zhong Su - "sdoc: exploring social wisdom for document enhancement in web mining" In *ACM* (2009), 395–404

[13] David Carmel, Haggai Roitman, and Elad Yom-Tov - "Social bookmark weighting for search and recommendation" In *VLDB* (2010), 761–775

[14] Pavel A Dmitriev, Nadav Eiron, Marcus Fontoura, and Eugene Shekita - "Using annotations in enterprise search" In *ACM* (2006), 811–817

[15] Kerstin Bischoff, Claudiu S Firan, Wolfgang Nejdl, and Raluca Paiu - "Can all tags be used for search?" In *ACM* (2008), 193–202

[16] David Carmel, Haggai Roitman, and Elad Yom-Tov - "Who tags the tags?: a framework for bookmark weighting" In *ACM* (2009), 1577–1580

[17] Sihem Amer Yahia, Michael Benedikt, Laks VS Lakshmanan, and Julia Stoyanovich - "Efficient network aware search in collaborative tagging sites" In *VLDB* (2008), 710–721

[18] Mohamed Reda Bouadjenek - "Infrastructure and Algorithms for Information Retrieval Based On Social Network Analysis/Mining" Thèse de Doctorat - *UVSQ* (2013)

[19] Marijn Koolen, Gabriella Kazai, and Nick Craswell - "Wikipedia pages as entry points for book search" In *ACM* (2009), 44–53

[20] Shihn-Yuarn Chen and Yi Zhang - "Improve web search ranking with social tagging" In *International Workshop on Mining Social Media* (2009)

[21] Yinghao Li, Wing Pong Robert Luk, Kei Shiu Edward Ho, and Fu Lai Korris Chung - "Improving weak ad-hoc queries using wikipedia asexternal corpus" In *ACM* (2007), 797–798

[22] Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su - "Optimizing web search using social annotations" In *ACM* (2007), 501–510

[23] Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, and Katsumi Tanaka - "Towards improving web search by utilizing social bookmarks". In *Web Engineering, Springer* (2007), 343–357

[24] Michael G Noll and Christoph Meinel - "Web search personalization via social bookmarking and tagging" In *Springer* (2007), 367–380

[25] Shengliang Xu, Shenghua Bao, Ben Fei, Zhong Su, and Yong Yu - "Exploring folksonomy for personalized search" In *ACM* (2008), 155–162

[26] David Vallet, Iván Cantador, and Joemon M Jose - "Personalizing web search with folksonomy-based user and document profiles" In Springer (2010), 420–431

[27] David Carmel, Naama Zwerdling, Ido Guy, Shila Ofek-Koifman, Nadav Har'El, Inbal Ronen, Erel Uziel, Sivan Yogev, and Sergey Chernov - "Personalized social search based on the user's social network" In *ACM* (2009), 1227–1236

[28] Matthias Bender, Tom Crecelius, Mouna Kacimi, Sebastian Michel, Thomas Neumann, Josiane Xavier Parreira, Ralf Schenkel, and Gerhard Weikum - "Exploiting social relations for query expansion and result ranking" In ICDEW 2008. *IEEE* (2008), 501-506

[29] Mohamed Reda Bouadjenek, Hakim Hacid, and Mokrane Bouzeghoub - "Sopra: A new social personalized ranking function for improving web search" In *ACM* (2013), 861–864

[30] Qihua Wang and Hongxia Jin - "Exploring online social activities for adaptive search personalization" In *ACM* (2010), 999–1008

[31] Xu Cheng, Cameron Dale, and Jiangchuan Liu - "Statistics and social network of youtube videos" In IWQoS 2008. *IEEE* (2008)

[32] Sally Jo Cunningham and David M Nichols - "How people find videos" In *ACM* (2008), 201–210

[33] Sergiu Chelaru, Claudia Orellana-Rodriguez, and Ismail Sengor Altingovde - "How useful is social feedback for learning to rank youtube videos?" In *WWW* (2014), 997–1025

[34] Bastian Karweg, Christian Hütter, and Klemens Böhm - "Evolving social search based on bookmarks and status messages from social networks" In *ACM* (2011), 1825–1834

[35] Ali Khodaei and Cyrus Shahabi - "Social-textual search and ranking" In *Proceedings of the First International Workshop on Crowdsourcing Web Search*, Lyon, France, (2012), 3–8

[36] P. Pantel, M. Gamon, O. Alonso, and K. Haas - "Social annotations: Utility and prediction modeling" In *ACM* (2012), 285–294

[37] G. Kazai and N. Milic-Frayling - "Effects of social approval votes on search performance" In *ITNG* (2009), 1554-1559

[38] Yoshiyuki Inagaki, Narayanan Sadagopan, Georges Dupret, Anlei Dong, Ciya Liao, Yi Chang, and Zhaohui Zheng - "Session based click features for recency ranking" In *AAAI* (2010), 1334-1339

[39] Ali Khodaei and Omar Alonso - "Temporally-aware signals for social search" In *SIGIR* (2012)

[40] Mohand Boughanem - "Modèles de langue pour la recherche d'information" available at : www.iro.umontreal.ca/~nie/IFT6255/modele_langue.pdf