Automatic relevant Source Discovery over the Internet based on user profile

Romain NOEL[1]

LITIS MIU - St Etienne du Rouvray
CASSIDIAN - Val-de-Reuil
Romain.Noel@cassidian.com

RÉSUMÉ. *La rapide croissance d'Internet ces dernières années a rendu difficile la découverte de nouvelles sources d'intérêt sur un sujet donné parmi l'important nombre de sources disponibles. Pour résoudre ce problème, nous présentons une approche permettant de fournir aux utilisateurs de nouvelles sources d'information pertinentes en exploitant leur besoin. Elle vise à combiner un robot d'indexation personnalisé et un système de filtrage collaboratif. Nous étudions l'intérêt d'exploiter le profil de l'utilisateur pendant le processus de découverte de sources afin d'éviter la collecte et l'indexation de tous les documents disponibles sur le Web. Nous discutons également des enjeux et du développement d'un tel système.*

ABSTRACT. *The enormous growth of the Web in recent years has made difficult the discovery of new sources of interest on a given topic, even thanks to an existing set of relevant sources. To address this problem, we introduce an approach to provide users with new relevant sources of information by exploiting their needs. It aims at combining a personalized crawler with a collaborative filtering system. We study the interest of exploiting the user profile during the crawling process in order to avoid collecting and indexing all available Web documents. We also discuss issues and the development of such a system.*

MOTS-CLÉS : *Modélisation de profil utilisateur, recherche d'information, retour utilisateur, exploration ciblée.*

KEYWORDS: *User modeling, information retrieval, ranking, relevance feedback, crawling.*

1. Supervised by : Nicolas MALANDAIN, Alexandre PAUCHET, Bruno GRILHERES, Laurent VERCOUTER

JJC 2012, pp. 401–406, Bordeaux, 21-23 mars 2012

1. Introduction

The explosive growth and popularity of the world-wide-web has involved a huge amount of information sources available on Internet. Automatic discovery of targeted information becomes a complex task. Therefore, search engines and crawlers have to focus on the most relevant sources. A relevant source can be defined as a web site or a part of a web site providing a set of relevant web pages according to user needs. Collaborative recommender systems can also emphasize relevant information for user individual needs. Resources of these systems are limited to knowledge already discovered. Contrary to crawlers, they are not able to explore the Web. We plan to combine a focused crawling system with a collaborative recommender system in order to benefit from each of their advantages. This approach can address the relevant source discovery and the monitoring of site mining problems.

The Section 2 introduces works related to personalized Information Retrieval (IR) and focused crawling. In Section 3, we propose a new relevant source discovery system based on user profile. In Section 4, a methodology for evaluating the proposed system is discussed. Finally, we conclude the paper in Section 5.

2. Related works

Personalized IR systems. A personalized search usually exploits a user profile modeling within a search engine. The user modeling, based on the user profile construction and representation, has a key role in the efficiency of personalized search. A user profile construction can be implicit or explicit. Implicit approaches automatically capture user interests. The technique most widely used consists in extracting information from user's search history such as query log or displayed Web pages (Tan *et al.,* 2006)(Liu *et al.,* 2004). The explicit construction involves that the user has to be proactive through an adequate user interface (Wærn, 2004). The user can communicate his preferences and interests to the system, by providing a set of relevant documents or by filling in questionnaires. The weakness of this approach is that the construction of the user profile is based on the user's willingness.

The user profile representations are usually based on vector models or bag of weighted-keywords (Salton *et al.,* 1975). Some approaches use ontologies to have a more structured user profile representation and to improve its quality. An ontology formally represents knowledge as a set of concepts. Some works use an ontology to determine search context and user's interests (i.e. Yago (Calegari *et al.,* 2010), DBPedia[1] (Bizer *et al.,* 2009) , ODP (Chirita *et al.,* 2005)).

Several approaches exploit a user profile in a IR system before or after the search engine process. The user profile can participate in a distinct re-ranking activity in order to increase the precision of the ordering process. The need representation can also be affected by the user profile during a pre-processing of the user query. User needs are

1. http ://dbpedia.org/About

mainly represented by queries. This approach uses the profile to transform user queries by adding or changing some keywords to improve the user needs.

Those approaches can improve the efficiency of search engines. However, those personalized search engines work on a set of indexed Web pages. This set of Web pages is collected from the whole Web without considering user needs. A strongly related topic is the "Focused Crawling" where techniques are used to reduce the scope of search engines which work beyond the crawling system.

Focused crawling. Crawlers explore Web pages from a URL queue. The URLs are converted into plain text to extract the contained links. Those links are added to the URL queue in order to crawl new Web pages. According to accepting rules, collected Web pages are then indexed. Focused crawling system exploits additional information : the hyper-textual information from Web pages such as anchors, text surrounding the links... This information is used to predict if the page can be relevant according to user interests. Focused crawlers analyze URL path (Chi *et al.,* 2001) or anchor text which provides information about page content. According to the topical locality phenomenon (Davison, 2000), Web pages on a specific topic are connected one to another. Focused crawlers are built on this phenomenon and crawl clusters of pages each time they find an interesting page. The two famous algorithms, PageRank (Page *et al.,* 1999) and HITS (Kleinberg, 1999), based on the structure of links, are often used to assign a hypertextual-based rank to seeds[2] to be explored. Chakrabarti et al. (Chakrabarti *et al.,* 1999) proposed a focused crawling system which makes use both of a classifier to evaluate the relevance of hypertext documents according to a specific topic, and of a distiller to identify hypertext nodes considered as good access points to other relevant sets of pages to collect.
In the literature, other approaches are related to the Semantic Web (Ehrig *et al.,* 2003) or use reinforcement learning (Rennie *et al.,* 1999) to improve the efficiency of crawlers. Despite the use of several techniques to explore only potential relevant sources, estimating the reliability of a source is difficult. On the other hand, some systems prefer to use traditional crawlers and search engines associated with a collaborative filtering system to fix the reliability problem with user feedback.

Personalized IR systems work on collection of documents whose size limit their efficiency. Estimate relevance of sources is a difficult task for crawlers. Considering those limitations, we propose to include personalized IR methods and user feedback techniques in a focused crawler to improve the discovery of relevant sources.

3. Proposed approach

The aim of our approach is to automatically discover relevant sources based on user needs. We propose a system composed of a crawler and a collaborative filtering recommender system. First, we plan to construct a model of user profile which can be used both by the crawler and the collaborative system. Next, we propose to use a

2. Sources used as starting URLs for the crawler

focused crawler to discover relevant sources before the indexing process. Finally, we aim at integrating a collaborative filtering system. The goal is to consider feedbacks on collected Web pages to enhance the source discovery process with relevant ones.

Features of the user profile. Our approach addresses this use case : a user has a set of URLs on a specific topic and he needs to find new relevant sources on this topic. Therefore, the construction of the user profile is based on a set of Web pages explicitly ordered by relevance and reliability. Zemanta[3] service is used to extract, for each ordered page, DBpedia concepts (Bizer *et al.,* 2009) and construct a weighted-concept vector by considering both the frequency of each concept into the set of Web pages, and also the rank of the Web page in the ordered list. This vector represents the user interests. The user can modify the list of concepts by re-weighting concepts or by removing irrelevant concepts.
The APML[4] (Attention Profiling Mark-up Language (Vuorikari, 2008)) allows to model user interest by compressing user data into a portable file format containing a description of ranked interests. This modeling methodology is particularly appropriate in our case since APML represents user profile with a list of concepts and sources(links). Our first contribution is to construct an APML ontology in order to use this model of user profile both by the crawler and by the collaborative system. Moreover, we plan to extend features of this ontology by adding information about the reliability of a source and information needed by the crawler or/and the collaborative system.

Discovery task. The set of URLs of the sources provided by the user are usually used as seeds for the crawler. Our approach consists of discovering relevant sources of interests before the collecting task. The crawler's seeds are not only the URLs provided by the user but also new URLs provided by a discovery process. As shown in figure 1, a temporary queue is added in order to receive those new URLs. The discovery process is composed of several task based on the user's URL list and on the concept list in order to find new relevant sources :
- Outgoing links : link extraction to capture outgoing links on user provided seeds.
- Back-links : URLs pointing to user provided seeds.
- Co-referenced URLs : social networks such as Twitter[5], and social bookmarking including Delicious[6] provide discussion topic pages. If user's URLs are mentioned, other relevant links mentioned into those pages can be extracted using API of Social Networks or "folksonomy" (Sinclair *et al.,* 2008) (Mika, 2005).
- Meta-search : concepts representing the user interest are exploited to create relevant queries submitted to famous search engine (Google, Yahoo !, Microsoft Bing).
- DBpedia links : data on DBpedia concepts are composed of interest links.
- Indexed URLs : URLs from similar context of crawl by users sharing common needs.

The goal is to improve the seeds list by adding new relevant URLs and to re-organize this list. The traditional URL queue is replaced by a ranked URL queue, in

3. http ://www.zemanta.com/
4. http ://www.apml.areyoupayingattention.com/
5. http ://www.twitter.com
6. http ://www.delicious.com

figure 1 (gray process), to crawl most relevant pages first. The system ranks the list of discovered seeds (Cho *et al.,* 1998), considering the type and the reliability of each seed. The system discovered seed is ordered according to the used ranked resource (seed or/and concept set).

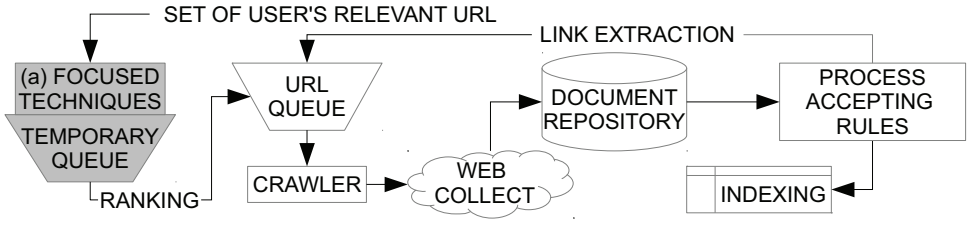

Figure 1. *Architecture of our crawler*

Document annotation and User feedback. A crawled document is annotated to preserve the context search and the user needs that have led to its collect. The system extracts concepts from crawled documents. The relevance of the seeds, the extracted concepts and their frequency of occurrence are used to annotate a document.
To present a ranked list of relevant documents to the user, the system has to consider both the semantic graph of each document and its *context graph*. A *context graph* is a representation of how a document can be accessed from the Web (Diligenti *et al.,* 2000). The system works only on a graph of crawled documents. This context graph represents how a document has been found from a list of seeds. The relevance of a crawled document is calculated considering its path in this Web sub-graph : cluster of pages are deduced from the distance between the document and seeds and the number of seeds pointing at the document. This sub-graph represents a distorted view of the Web. We also work on a semantic graph. The system extracts concepts from the content of those crawled documents. A semantic graph is constructed thanks to hierarchical relations into the DBpedia ontology. The system measures the distance between document concepts and the user's concepts. We consider the relevance of concepts establish during the user profile construction, and the occurrence rate of a concept into a crawled document. Considering those two graphs, the seeds rank can be propagated to discovered sources in order to rank our crawled documents by relevance.
A feedback mechanism allows the user to express if a document is not interesting. This information is added to the document metadata. The ranking process exploits this information to rank seeds provided by the *indexed URLs* technique.

4. Conclusions

To solve the problem of discovering new relevant sources, we propose a system which combine personalized IR methods with feedback techniques and a focused crawler. We plan to construct a model of user profile which can be exploited during a discovery task to suit user needs. We will evaluate the efficiency of the discovery task by analyzing the rate of relevant documents among the set of crawled documents. A user experiment is planned in order to assess the efficiency of our system.

5. Bibliographie

Bizer C., Lehmann J., Kobilarov G., Auer S., Becker C., Cyganiak R., Hellmann S., « DBpedia-A crystallization point for the Web of Data », *Web Semantics : Science, Services and Agents on the World Wide Web*, vol. 7, n° 3, p. 154-165, 2009.

Calegari S., Pasi G., « Ontology-Based Information Behaviour to Improve Web Search », *Future Internet*, vol. 2, n° 4, p. 533-558, 2010.

Chakrabarti S., Van den Berg M., Dom B., « Focused crawling : a new approach to topic-specific Web resource discovery », *Computer Networks*, vol. 31, n° 11-16, p. 1623-1640, 1999.

Chi E., Pirolli P., Chen K., Pitkow J., « Using information scent to model user information needs and actions and the Web », *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, p. 490-497, 2001.

Chirita P., Nejdl W., Paiu R., Kohlschütter C., « Using ODP metadata to personalize search », *Research and development in information retrieval*, ACM, p. 178-185, 2005.

Cho J., Garcia-Molina H., Page L., « Efficient crawling through URL ordering », *Computer Networks and ISDN Systems*, vol. 30, n° 1-7, p. 161-172, 1998.

Davison B., « Topical locality in the Web », *Research and development in information retrieval*, ACM, p. 272-279, 2000.

Diligenti M., Coetzee F., Lawrence S., Giles C., Gori M., « Focused crawling using context graphs », *Very large data bases conference*, p. 527-534, 2000.

Ehrig M., Maedche A., « Ontology-focused crawling of Web documents », *Applied computing*, ACM, p. 1174-1178, 2003.

Kleinberg J., « Authoritative sources in a hyperlinked environment », *Journal of the ACM (JACM)*, vol. 46, n° 5, p. 604-632, 1999.

Liu F., Yu C., Meng W., « Personalized web search for improving retrieval effectiveness », *IEEE Transactions on knowledge and data engineering*p. 28-40, 2004.

Mika P., « Ontologies are us : A unified model of social networks and semantics », *The Semantic Web-ISWC 2005*p. 522-536, 2005.

Page L., Brin S., Motwani R., Winograd T., « The PageRank citation ranking : Bringing order to the web. », 1999.

Rennie J., McCallum A., « Using reinforcement learning to spider the web efficiently », *Machine learning international workshop*, p. 335-343, 1999.

Salton G., Wong A., Yang C., « A vector space model for automatic indexing », *Communications of the ACM*, vol. 18, n° 11, p. 613-620, 1975.

Sinclair J., Cardew-Hall M., « The folksonomy tag cloud : when is it useful ? », *Journal of Information Science*, vol. 34, n° 1, p. 15, 2008.

Tan B., Shen X., Zhai C., « Mining long-term search history to improve search accuracy », *International conference on Knowledge discovery and data mining*, ACM, p. 718-723, 2006.

Vuorikari R., « Consolidating collections of learning resources using APML », *Mashup Personal Learning Environments (MUPPLE08)*, vol. 388, p. 14-17, 2008.

Wærn A., « User involvement in automatic filtering : An experimental study », *User Modeling and User-Adapted Interaction*, vol. 14, n° 2, p. 201-237, 2004.