Pakistan Journal of Information and Technology 2 (2): 96-103, 2003
ISSN 1682-6027

A Standard Framework for Personalization Via Ontology-Based Query Expansion

Jinan Fiaidhi, Sabah Mohammed, [1]Jihad Jaam and [1]Ahmad Hasnah
Department of Computer Science, Lakehead University,
Thunder Bay, Ontario P7B 5E1, Canada
[1]Department of Computer Science, Qatar University, P.O. Box 2713, Doha, Qatar

Abstract: As the number of available Web pages grows, users experience increasing difficulty finding documents relevant to their interests. One of the underlying reasons for this is that most search engines find matches based on keywords, regardless of their meanings. To provide the user with more useful information, we need a system that includes information about the conceptual frame of the queries as well as its keywords. Moreover, web searching lack standard marks, standard ways of interacting with users, benchmarks tests and even a standard terminology, thus presenting opportunities for developing application specific search hosting. This article develops a standard framework for designing personalized search engines. The framework composed of three plugs-in components: training, spreading and filtering, which can be attached to any search engine.

Key words: Personalization, search engine, query expansion

Introduction

General web search is performed predominantly through text/keywords queries to search engines. Searching the explosive content on the Internet merely on general search is certainly not a very smart idea. The main issues of keyword-based query model lie in the difficulty of query formulation and the inherent word ambiguity in natural language. The problem is best illustrated through the scenario of information search on the Web, where the queries are usually of two words long and a large number of "hit" documents are returned to the user. Part of the reason comes from the inherent ambiguity of word in natural language. Another part is the difference of interpretation for a query. That is, given the same query expression by different users, the information inquired could range from various perspectives. Certainly, matching the information need of the Internet users with the content on the Web requires modeling of the user needs. Such type of personalized searching can be addressed under the umbrella of Search Hosting. Search hosting deals with techniques for tailoring a user's query with the Web information space based on personalized information.

The work on personalizing web searching started with various practical attempts to construct a personalized search engines such as Web Watcher (Joachims *et al.,* 1997); WebMate (Chen and Sycara, 1998); Amolthoea (Moukas, 1996) and Alipes (Widyantoro *et al.,* 1999). All such search engines attempt to automatically filter web pages on behalf of the user, based on his/her

previous monitored keywords profile. Another practical direction that contributed to personalization of web search is classification. It is an attempt to organize information by classifying or categorizing documents into the best matching category in a predefined set of categories. There are several attempts in this direction (Has *et al.,* 1999; Gover *et al.,* 1999) where aim is to anticipate the classifications of the web pages into document type according to the pages structural characteristics. Ontologies, on the other hand, is a newly attempt to structure searching information through the use of graph of concepts. Recently several search engines attempts to include ontology in their searching mechanism such as OntoSeek (Guarino *et al.,* 1999); Telltale (Chowder and Nicholas, 1996); SHOE (Heflin *et al.,* 1999) and OBIWAN (Chaffee and Gauch, 2000). All such attempts and approaches lucks a uniform framework for personalizing web search which takes into account the topic or concepts related to the user query.

Related research work

Previously researchers working under the umbrella of search hosting have focused their efforts on page ranking (McGill *et al.,* 1979); automatic query expansion (Crouch and Yang, 1992); relevance feedback (Salton and Buckley, 1990) and other hybrid techniques (Harman, 1996; Greenberg and Garber 1991) to help the user formulate what information is really needed and clarify their query ambiguity. The *PageRank* algorithm was proposed to exploit the linkage structure of the web to compute global "importance" scores that can be used to influence the ranking of search results so the number of query results can be limited to a manageable size. Since different users may have preference for different web pages, the query results should also encompass this notion of importance. A "personalized view" of the web can be achieved by modifying the PageRank algorithm with a given a personalization vector (or preference vector) u drawn from hub set H, (details will be described later) and this personalized view is represented by a personalized PageRank vector (PPV) v. However, computing a PPV naively using a fixed-point iteration requires multiple scans of the web graph, which makes it too expensive to compute online in response to a user query. On the other hand, there are 2^n different personalization vectors, (n is the length of the personalization vector), which makes it too expensive to store offline.

Automatic query expansion enhances web search by adding new words to these queries via blind feedback, without any input from the user. The promises of such retrieval are great. However, the implementation of automatic query expansion has not proven as useful as originally desired. Nevertheless, a considerable amount of research has gone on in the development of automatically derived thesauri and query expansion techniques. They can be divided into various categories again depending on the methods used. The relationship between terms within a document and in the wider collection lies at the heart of such systems.

Relevance feedback, however, is a semi-automatic procedure, wherein the information system formulates new queries based on user input. In essence, the process looks something like this: The user formulates an initial query, which results in a primary retrieval set. The user then selects from this list documents that they determine are relevant to their information need,

which are in turn used by the system to re-weight, expand and/or reformulate a new query for searching. The simplest example of such a system would offer the user the ability to locate relevant documents and select "more like this." Although query reformulation and query expansion are in practice different, the same information seeking theories underlie both processes in their ultimate implementation. Query expansion using relevance feedback methods can take on various forms, depending on the theoretical model employed and method used for expanding the query. Salton's vector space model is most often used, although past work has been done using the probabilistic retrieval model (Van Rijsbergen, 1979) and Boolean systems (Salton *et al.,* 1990). Methods for expansion differ depending on the number and type of terms drawn from relevant documents. The first of these doesn't actually expand the query at all, instead simply re-weighting the terms in the original query to more appropriately reflect the chosen relevant document or documents. On the other end of the scale is full query expansion, which expands the query with all the terms in the chosen relevant documents. In between these extremes are a host of partial query expansion techniques, which often times select either the most frequently occurring terms or the most highly weighted terms in the relevant documents. According to Salton and Buckley (1990), the best overall relevance feedback method is what is known as the "Ide dec-hi" method (Ide, 1971). In this approach, all of the designated relevant documents and the highest retrieved *non*-relevant document are used in reformulating the query. The latter is chosen as a "definitive point in the vector space from which the new feedback query is removed." Using the "Ide dec-hi" method, experiments showed up to 160% improvement over non-expanded queries. Problems exist, however, in actually implementing relevance feedback systems. Research has shown that users prefer not to be bothered with manually having to offer feedback to the system. Among the most successful hybrid techniques are the "interactive query expansion" and the "pseudo-relevance feedback". Iterative query expansion connects the relevance feedback with thesauri (Efthimiadis, 1996). After submitting an initial query, the system with iterative query expansion presents users with a list of associative or related terms drawn from either a handcrafted thesaurus or derived automatically from the collection or the retrieval set. Although proven to be useful, more research is needed to compare these results with traditional document-based relevance feedback mechanisms.

The problems associated with relevant feedback let recent research to concentrate on what is known as pseudo-relevance feedback (Xu and Croft, 1996). Such systems retrieve a number of documents after an initial query search. Assuming that the top-n documents are relevant, the system takes the terms from these documents and expands the query. Of course, such systems depend largely on the effectiveness of the system to choose relevant documents in the first place. If these documents are not relevant, non-relevant search terms will be added to the query automatically, thus degrading effectiveness.

The main point to make against the previously used methodologies for search hosting is that they pay attention only to shared vocabulary/keywords and ignores the other personalization semantics related to document searching where document belongs to a certain category, keywords describe a category, a keyword can be a synonym of another and one category is a

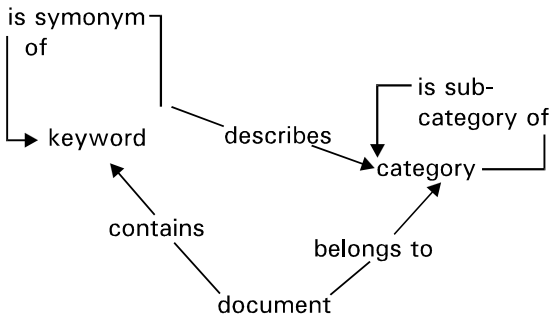

Fig. 1: Document searching semantics

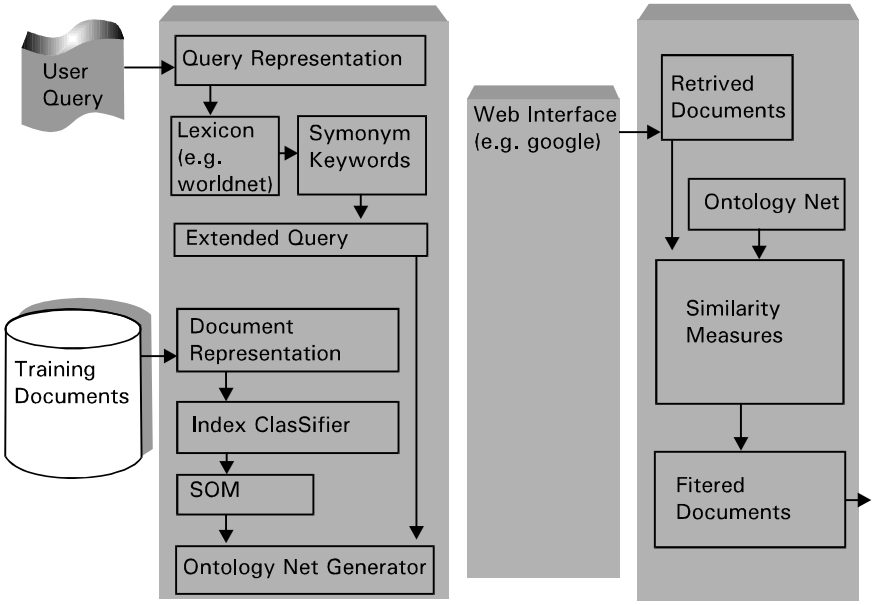

Fig. 2: General framework for search hosting

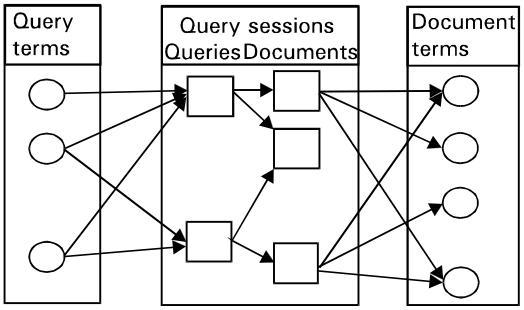

Fig. 3: Query sessions, query terms and document terms

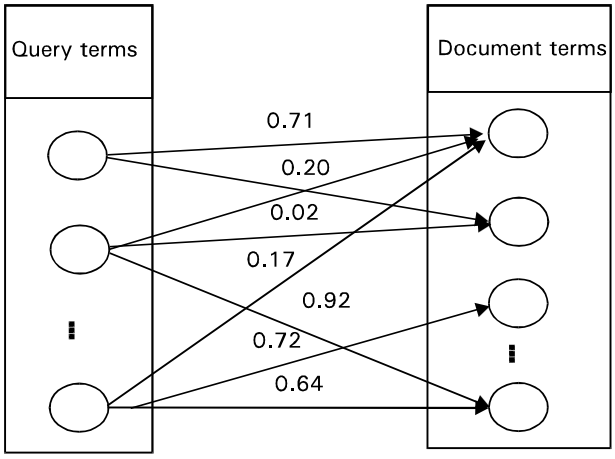

Fig. 4: Ontology Net

subcategory of another. Fig. 1 visualizes the other factors effecting document searching.

One needs a framework that takes into account such semantics relationships to guide hosted search through a machine learning algorithms to find previously unknown knowledge online.

Why ontology-based query expansion for search hosing

A lack of vocabulary compatibility between user and the information system often impairs searches. This may be due to lack of or differences in expertise and the representation of the document semantics domains. For this purpose researchers started to understand the importance of these factors which is generally called the search ontology. In this direction, the concepts, relationships and rules related to the document searching domain can already be considered an ontology. According to Chaffee, an ontology is an arrangement of concepts that represents a view of the world(Chaffee and Gauch, 2000) that can be used to structure information. Ontologies can be built by specifying the semantic relationships between the terms in a lexicon.

Ontology attracts attentions across many fields in computer science recently. The term ontology originates from philosophy and its current usages in computer science (first introduced by people in AI) is far from its philosophical origin. There exists no consensus definition about ontology. One most cited is "Ontology is an explicit representation of a conceptualization, the conceptualization includes a set of concepts, their definition and inter-relationships"(Gruninger and Lee, 2002). In many cases, the term ontology is another name denoting the result of familiar activities like conceptual analysis and domain modeling. The roles of ontology vary from knowledge management to semantic interoperability. One important reason for that ontology attracts so many attentions recently is the semantic web, since ontology is considered as the key enabler of semantic web. Currently the semantic web community is working on standards for the representation and exchange of ontologies via the Internet. One of the most prominent

approaches is the Ontology Inference Layer (OIL) (Klein Fensel *et al.,* 2000).

Towards a standard framework for personalization

Fig. 2 illustrates the general idea of designing a framework for search hosting. This framework can be used to generate architectures for personalized search engines. It consists of three phases: Training, Spreading and Filtering. The initial user query will be first expanded through the use of a lexicon (e.g. WorldNet (Fellbaum, 1997)). The training phase comprises of an index classifier and a self-organizing map. During the index classifier training, a fixed number of sample documents for each concept are collected and merged and the resulting super-documents are preprocessed and indexed using a suitable ranking method (Drori, 2002). The result of this stage is a relevant document ranking. The self-organizing map uses these rankings as well as the user searching activities log to train the ontology net (Kangas, 1994). The spreading phase takes inputs from the expanded query and from the ontology net to produce new query. This new query is then feed to a Web interface (e.g. Google (Page, 2002)) and the resulted search is feed to a searching filter where it can be compared with the ontology net for similarity. Only those proven to be highly similar documents can be released for browsing. Measures like Recall and Precision (Raghavan *et al.,* 1989) or Fuzzy Ontology Pruning (Widyantoro *et al.,* 1999) can be used to in evaluating the similarity between the ontology net and the retrieved documents. All these components can be designed as a Plug-in (Gran and Scheller, 2000) interfaces which can be attached to web browsing interface.

The ontology net generator tries during the query sessions to bridge the gap between the query space and the document space. Fig. 3 shows how correlations between the query terms and document terms can be established through the query sessions. In general, we assume that the terms in a query are correlated to the terms in the documents that the user clicked on. If there is at least one path between one query term and one document term, a link is created between them. By analyzing a large numbers of such links, we can construct an ontology net for the correlations between the terms in these two spaces (Fig. 4).

Searching and mining the Web, as well as analyzing user behavior while using the Web, are exciting areas of research. We have reviewed some recent results in this area that are not only technically satisfying, but also have the potential to significantly impact searching, browsing and collaboration among Web users. Research attempts like page ranking, automatic query expansion, relevance feedback as well as many other hybrid methods (e.g. iterative query expansion) and dedicated search engines(e.WebLog, WebSQL) has cited with different goals and performance. A plethora of research centers have been setup world wide to carry out research activities related to Web data analysis. None of these centers uses standard techniques for query expansion. This article present a standard framework for designing personalized search engine where we can host our future searches. Three important components has been identified for this purpose: Training, Spreading and Filtering. Such components can be designed as Plugs-in components to be attached to any general web interfacing engine.

References

Chen, L. and K. Sycara, 1998. A Personal Agent for browsing and Searching, In Proceedings 2nd Intl. Conference on Autonomous Agents, N.Y.

Chowder, G. and G. Nicholas, 1996. Reseource Selection in Café: An Architecture for Networked Information Retrieval, In Proceedings SIGIR'96, Workshop on Networked Information Retrieval, Zurich.

Chaffee, J. and S. Gauch, 2000. Personal Ontologies for Web Navigation, Conference on Information and Knowledge Management, 9th Intl. Conf. On Information and Knowledge Management, Virginia, USA.

Crouch, C.J. and B. Yang, 1992. Experiments in automatic statistical thesaurus construction. In Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, ed. N. Belkin, P. Ingwersen and A. M. Pesjtersen: New York: ACM Press, pp: 21-24.

Drori, O., 2002. Algorithm for Document Ranking: idea and simulation, ACM Proceedings of the 14th Intl. Conference on Software Engineering and Knowledge Engineering, Ischia, Italy.

Efthimiadis, E., 1996. Query expansion. Annual Review of Information Systems and Technology (ARIST), 31: 121-187.

Fellbaum, C., 1997. WordNet: An Electronic Lexical Database, MIT Press.

Gover, N., M. Lalmas and N. Fuhr, 1999. A Probablistic Description-Oriented Approach for Categorizing Web Documents, In Proceedings 8th Intl. Conference on Information and Knowledge Management.

Guarino, N. *et al*., 1991. OntoSeek: Content-Based Access to the Web, IEEE Intelligent Systems.

Greenberg, I. and L. Garber, 1991. Searching for New Search Technologies, IEEE Computer.

Gruninger, M. and J. Lee, 2002. Ontology: Applications and Design. Communications of the ACM, 45: 39-41.

Gran, C. and A. Scheller, 2000. From Proven Office Technologies to Intelligent Multimedia, IEEE Intl. Conference on Multimedia and Expo 2000, N.Y.

Has, W., 1999. Classification Algorithms for NetNews Articles, In Proceedings 8th Intl Conference on Information and Knowledge Management.

Heflin, J., J. Hendler and S. Luke, 1999. SHOE: A Knowledge Representation Language for Internet Applications. Technical Report, CS-TR-4078 (UMIACS TR-99-71), Dept. Computer Science, Uni. Maryland at College Park.

Harman, D., 1996. Towards interactive query expansion. In Proceedings of the Seventh ACM Conference on Hypertext, Washington, DC.

Ide, E., 1971. New experiments in relevance feedback. In G. Salton. The SMART Retrieval System: Experiments in automatic document processing. Englewood Cliffs, NJ: Prentice-Hall.

Joachims, T., D. Freitag and T. Mitchell, 1997. WebWatcher: A Tour Guide for WWW, In: Proceedings IJCA'97.

Klein, M., D. Fensel, F. van Harmelen and I. Horrocks: The relation between ontologies and schema-languages: Translating OIL-specifications in XML-Schema. In: Proceedings of the Workshop on Applications of Ontologies and Problem-solving Methods, 14th European Conference on Artificial Intelligence ECAI'00, Berkub, Germany, pp: 20-25.

Kangas, J., 1994. On the Analysis of Pattern Sequences by Self-Organizing Maps. Doctorate Thesis, Helsinki Univ. of Tech., Lab. of Comp. and Inf. Science, SF-02150. Espoo, Finland.

Moukas, A., 1996. Amathaea: Information Discovery and Filtering using a MultiAgent evoving ecosystem, In Proceedings 1st Intl. Conference on Practical Applications of Intelligent Agents and MultiAgent Technology, London.

McGill, M., M. Koll and T. Nomault, 1979. An: Evaluation of Factors Affecting Document Ranking by information Retrieval Systems. Report, School of Information Studies, Syracuse University, Syracuse, New York

Page, L., 2002. Google History. Google Inc. http://www.google.com/corporate/history.html.

Raghavan, V., G. Jung and P. Bollman, 1989. A critical investigation of recall and precision as measures of retrieval system performance. ACM Transactions on Information Systems, 7: 205-229.

Salton, G. and C. Buckley, 1990. Improving retrieval performance by relevance feedback. J. American Soc. Inform. Sci., 41: 288-297.

van Rijsbergen, C.J., 1979. Information Retrieval. London: Butterworths. Salton, G., E. Voorhees and E. Fox. 1984. A comparison of two methods for Boolean query relevancy feedback. Information Processing and Management, 20: 637-651.

Widyantoro, D., T. Lorger and J. Yen, 1999. An adaptive algorithm for learning changes in user interests. In: Proceedings 8th Intl. Conference on Information and Knowledge Management.

Xu, J. and B. Croft, 1996. Query expansion using local and global document analysis. In Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, Ed. H. P. Frei and P. Schauble: New York: ACM Press, pp: 4-11.