International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:2, No:6, 2008

Information Retrieval in Domain Specific Search Engine with Machine Learning Approaches

Shilpy Sharma

***Abstract*—**As the web continues to grow exponentially, the idea of crawling the entire web on a regular basis becomes less and less feasible, so the need to include information on specific domain, *domain-specific search engines was* proposed. As more information becomes available on the World Wide Web, it becomes more difficult to provide effective search tools for information access. Today, people access web information through two main kinds of search interfaces: *Browsers* (clicking and following hyperlinks) and Q*uery Engines* (queries in the form of a set of keywords showing the topic of interest) [2]. Better support is needed for expressing one's information need and returning high quality search results by web search tools. There appears to be a need for systems that do reasoning under uncertainty and are flexible enough to recover from the contradictions, inconsistencies, and irregularities that such reasoning involves. In a *multi-view* problem, the features of the domain can be partitioned into disjoint subsets (*views*) that are sufficient to learn the target concept. Semi-supervised, multi-view algorithms, which reduce the amount of labeled data required for learning, rely on the assumptions that the views are *compatible* and *uncorrelated.* This paper describes the use of semi-structured machine learning approach with Active learning for the "Domain Specific Search Engines". A domain-specific search engine is "An information access system that allows access to all the information on the web that is relevant to a particular domain. The proposed work shows that with the help of this approach relevant data can be extracted with the minimum queries fired by the user. It requires small number of labeled data and pool of unlabelled data on which the learning algorithm is applied to extract the required data.

***Keywords*—**Search engines; machine learning, Information retrieval, Active logic.

I. Introduction

THE paper is organized as follows: Section II specifies WWW, Search engines are discussed in Section III, section IV and section V covers various types of machine learning approaches and active learning approach. In section VI, problems with the existing system have been discussed and section VII covers the proposed remedies as per the limitations of the existing system.

II. WWW

World Wide Web is a vast source of information organized in the form of a large distributed hypertext system. Every Hypertext link on every web page in the world contains one of the URLs. When you click on a link of any kind on a Web page, you send a request to retrieve the unique document on some computer in the world that is uniquely identified by that URL. URLs are like addresses of web page. Information Retrieval can be done in two ways as shown in the Fig. 1.

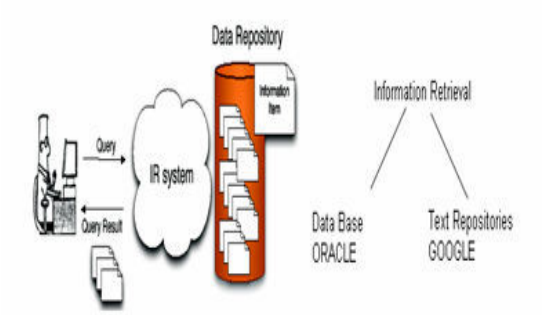

Fig. 1 Information Retrieval

III. Search Engine

A ***search engine*** is a program [1] that can search the Web on a specific topic. By typing in a word or phrase (known as a keyword), the search engine will produce pages of links on that topic. The more relevant links are at the top of the list, but that is not always true. Information retrieval system works as shown in the Fig. 2.

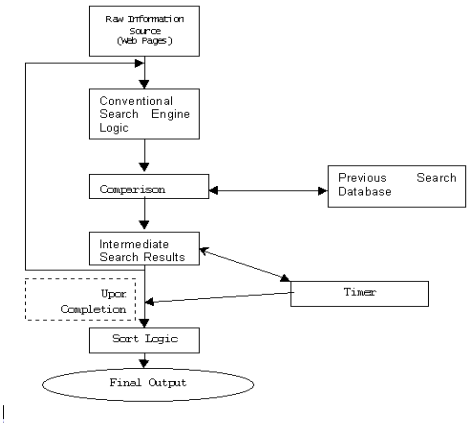

Fig. 2 Information Retrieval Architecture

S. Sharma is with CSE-IT Department, Institute of Technology and Management, Gurgaon, India (e-mail: Sharma.shilpy@gmail.com).

We can classify search engines as *general-purpose search engines* and *domain-specific search engines.* Search engines come in three major flavors: Web crawlers, Web portals, Meta-Search engines. A search engine or a web crawler has several parts: Firstly, a *crawler* that traverses the web graph and downloads web documents; Secondly, an *indexer* that processes and indexes the downloaded documents; Thirdly, a *query manager* that handles the user query and returns relevant documents indexed in the database to the user. The Web is similar to a graph, in that links are like edges and web pages are like nodes. Crawler starts from a *seed* page and then uses the external links within it to attend to other pages. Web crawlers o retrieve Web pages and add them or their representations to a local repository. Web Content Analysis is done either by *Document-query similarity* which is important for identifying similarity of user query to documents or *Document-document similarity* which is important for finding similar documents in the document pool of a search engine. **The analysis Link Structure of Web is done by *ranking or Hub and Authority pages.* In the former *w*eb pages differ from each other in the number of in-links that they have. Page Rank Algorithm is used for the purpose and in later the importance of pages can be extracted from the link structure of web. In this approach two kinds of pages are identified from web** page links: *first* pages that are very important and authorities in a special topic, *second* pages that have great number of links to authority pages.

IV. Machine Learning

Machine learning is the study of how to make computers learn; the goal is to make computers improve their performance through experience. As shown in Fig. 3 computer is getting the input as a class of tasks and with the help of algorithm, the performance of the system can be improved.

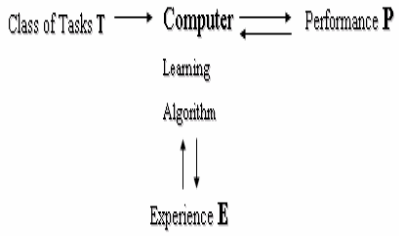

Fig. 3 Machine Learning

Machine learning approaches in Information Retrieval is the ability of a machine to improve its performance based on previous results [13]. Machine learning is an area of artificial intelligence concerned with the development of techniques, which allow computers to "learn". More specifically, machine learning is a method for creating computer programs by the analysis of data sets. The Machine Learning is of three types:

- *Supervised learning*: it is also called classification – is based on learning from a training data set.
- *Unsupervised Learning*: In this approach data exists but there is no knowledge of the correct answer of applying the model to the data. An unsupervised learner agent is neither told the correct action in each state, nor is an evaluation of action taken in each state.
- *Semi-supervised learning:* it is a goal-directed activity, which can be precisely evaluated either Learning from labeled and unlabeled documents, Reinforcement learning and Case Based Reasoning.

V. Active Learning

Active learning is to design and analyze learning algorithms that can effectively filter or choose the samples for which they predict the hypothesis.

VI. Limitations of Existing System

As the web continues to grow exponentially, the idea of crawling the entire web on a regular basis becomes less and less feasible. General-purpose search engines have certain limitations like *Querying Mechanism, Keyword Exact Matching, Low web Coverage Rate, Long result list with low relevancy to user query.* Due to these basic problems and also because of the need to include information on specific domain, *domain-specific search engines were* proposed. A domain-specific search engine is defined as: "an information access system that allows access to all the information on the web that is relevant to a particular domain". Search engines can be classified as general-purpose search engines and domain-specific search engines. *"Domain"* in domain-specific search engines can be specialized with two existing kinds of search engines, search engines which focus on ***specific document type*** such as resumes, homepages, movies, etc. And search engines focus on ***specific topic*** like computer science, climbing, sport, etc Domain –Specific search engines can be IBM- Focused Crawler, Cora Domain Specific Search Engine.

A. IBM- Focused Crawler

Focused crawling, as Soumen Chakrabati et al [7, 14] proposed a new approach to topic specific resource discovery at IBM's Almaden center. The focus topic in this system is represented by a set of example pages that is provided by a user to the system. In the system described there is a user-browsable topic taxonomy where the user can mark some of the documents as *good* and select them as the focus topic. The system has three main components: A *classifier* that makes judgments on the relevancy of crawled documents, and decides on following the links within pages. The classifier is an extended version of the Naïve Bayes classifier. The second component is a *distiller* that evaluates the centrality of each page to determine crawling priority of links within it. The distiller uses the bibliometric concepts of *hub* and *authority* pages as an approximate social judgment of web page quality. For each page it calculates the hub and authority scores of each web page with an extended version of the HITS algorithm. It tries to crawl the links within pages with the highest hub score first in order to find new authorities. The

third component of the system is a dynamic *crawler* that crawls the web according to a re-orderable priority queue. The system works in two phases: training and testing. In the training phase, the classifier is trained with some labeled data relevant to the focus topic. The training data set is acquired from existing taxonomy-like search engines (portal) such as Yahoo! and Open Directory Project.

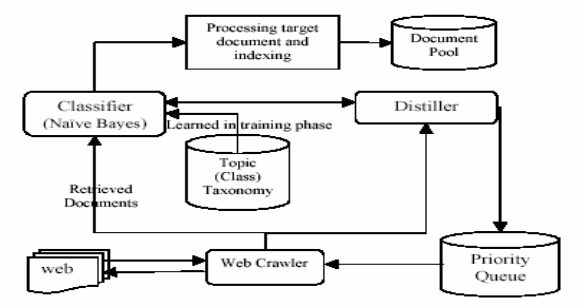

Fig. 3 IBM focused Crawler

Problems associated with this search engines are:

- ***Fixed model of classifier***, IBM focused crawler uses a fixed model of relevancy class as a classifier to evaluate topical relevancy of documents. A more adaptive classifier uses documents that are marked as relevant by the classifier to update the classifier. However, ensuring flexibility in the classifier without simultaneously corrupting the classifier is difficult.
- ***Does not model future rewards***. One major problem faced by this focused crawler is that it is does not learn that some sets of off-topic documents often lead reliably to highly relevant documents. In other words it does not model the future reward of links. For example, a home page of a computer science department is not relevant to "Reinforcement Learning", but links from that home page eventually lead to the home page of the "Machine Learning Laboratory" or the home page of a researcher where it may find valuable target pages.
- ***Lack of comparison of results***. The results reported in papers of this approach are not compared with results of human-maintained portals like the Open Directory Project. However, judgments on the quality of gathered pages in this approach are hard.
- ***Exemplary documents***. Representation of focus topic in the form of some high quality documents related to topic is sometimes hard for the user.

VII. Proposed Work

The problems that have been discussed in section 6 can be removed with the help of machine learning approaches along with the active learning approach. As discussed in the section 4 that machine learning can be of three types, by employing the semi structured machine-learning approach to the IBM focused crawlers the system performance can comparatively be improved. Semi-supervised learning is a goal-directed activity, which can be precisely evaluated. In the web context our training data is a small set of labeled documents. The label is document class, and our goal is to guess the label of an un-seen document. In this category we review learning from labeled and unlabeled documents. In some semi-supervised approaches, a learner agent learns from interaction with a dynamic environment. In these environments, providing a set of training data for the agent is very difficult or even impossible, because of the dynamics inherent in the environment and correspondingly huge number of states and actions. One requirement of this model is a measure of the goodness of action that the agent takes in a state.

A. Active Learner in Domain Specific Search Engine

To overcome the problems that occur in IBM, focused crawler, one technique is to replace the classifier with active learner. As discussed in section V, Active learner is capable of predicting the hypothesis [11]; it will generate the hypothesis with the experiences. Semi-supervised learning is used because it is very expensive to generate labeled data for every set. To use active learning approaches with semi-supervised learning to improve the efficiency of the system. Semi-supervised algorithms are used in text classification:

- Co-Training
- Semi-supervised EM
- Co-EM
- Co-EMT uses a multi-view active learning algorithm

Multi-view setting applies to learning problems that have a natural way to divide their features into subsets (*views*) each of which are *sufficient* to learn the target concept. Multi-view active learning maximizes the accuracy of the learned hypotheses while minimizing the amount of labeled training data.

The algorithm to improve the efficiency is discussed as follows:

Given

- *A learning problem with view*
- *A learning algorithm*
- *Set of "T" and "U" labeled and unlabeled examples*
- *No. of k iterations to be performed*
- *The number N of queries to be made*

- *CO-EMT for domain specific engine*
  - *Let iter be the number of iterations to be made*
  - *LOOP for k iterations*
  - *- use L, V1(T), and V2(T) to create classifiers h1 and h2*
  - *- FOR EACH class Ci DO*
  - *- let E1 and E2 be the e unlabeled examples on which h1and h2 make the most confident predictions for Ci(mean the relevant data according to the user query)*
  - *- label E1 and E2 according to h1and h2,respectively, and add them to T*

- *combine the prediction of h1 and h2*
- *this combination will be the combination of the required documents.*

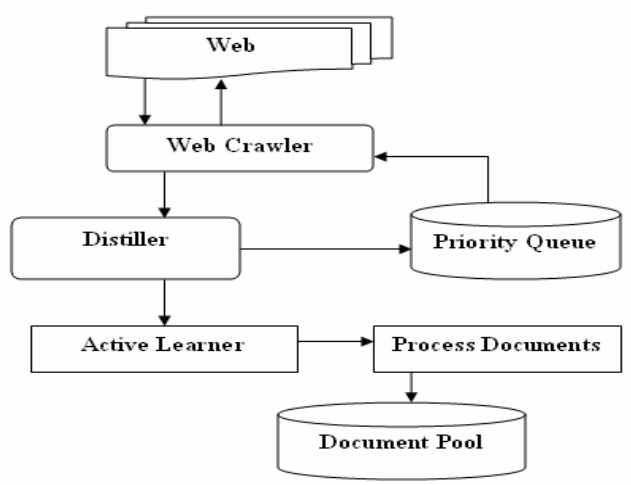

Fig. 5 Improvement in IBM focused Crawlers

The system has three main components: (1) Active *learner* makes judgments on the relevancy of crawled documents, the tasks that are being done by active learner: Firstly, it maintains the history and makes hypothesis based on them, Secondly it allows meta- reasoning to be done. (2) *Distiller* that evaluates the centrality of each page to determine crawling priority of links within it. The distiller uses the bibliometric concepts of *hub* and *authority* pages as an approximate social judgment of web page quality. For each page it calculates the hub and authority scores of each web page with an extended version of the HITS algorithm. (3) Dynamic *crawler* that crawls the web according to a re-orderable priority queue.

VIII. Conclusion

Search engine technology has gone through several evolutions and finally reached the point where Artificial Intelligence can offer tremendous help. We have reviewed this evolution from the beginning up to now and surveyed several different techniques that have been developed to improve search engine functionality. In particular we highlighted some machine learning approaches to information retrieval on the web and concentrated on topic-specific search engines. Finally, we proposed an information integration environment based on active learning. Our approach uses current technology in a better manner to provide appropriate results.

References

[1] LookOff E-book, Engine Basics, http://www.lookoff.com/tactics/engines.php3 , Oct 24 2000.
[2] M. Jaczynski, B. Trousse, Broadway: A Case-Based System for Cooperative Information Browsing on the World-Wide-web, Collaboration between Human and Artificial Societies, pp. 264-283, 1999.
[3] Internet Fact and State, http://optistreams.com/factsandstats15.htm
[4] The Censorware Project, http://www.censorware.org/web_size, Jan. 26, 1999
[5] S. Lawrence and C.L. Giles, Searching the World Wide Web, Science 80:98-100, 1998.
[6] S. Lawrence and C.L. Giles, Accessibility of Information on the Web, Nature 400:107-109,1999.
[7] S. Chakrabarti, Data mining for hypertext: A tutorial survey, SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM 1(2): 1-11, 2000.
[8] L. Page, S. Brin, The anatomy of a large-scale hypertext web search engine, Proceeding of the seventh International World Wide Web Conference, 1998.
[9] S. Mizzaro, Relevance: The whole history, Journal of the American Society for Information Science, 48(9): 810-832, 1997.
[10] S. Lawrence, Context in web Search, IEEE Data Engineering Bulletin,
[11] Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. Proc. of the Conference on Computational Learning Theory (pp. 92–100).
[12] Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. Proc. of the Empirical NLP and Very Large Corpora Conference (pp. 100–110). de Sa, V., & Ballard, D. (1998).
[13] T. M. Mitchell, Machine Learning, New York: McGraw-Hill, 1997.
[14] S. Chakrabarti, M. van der Berg, and B. Dom, Focused crawling: a new approach to topic-specific web resource discovery, Proceeding of the 8th International World Wide Web Conference (WWW8), 1999.