An Improved Search Engine by Semantic Web Using Ontology

P. Hema Priya[1], R. RangaRaj[2]

[1]Research Scholar, Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, India

[2]Head, Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, India

Abstract: ***The internet contains vast amount of information that the search engines are able to provide search results that are based on page ranks. But the search results are not related to one particular user's environment. The main objective of this paper involves with search engine and search engine optimization methods. Here a new technique called as ontology search logs is introduced, which will be used for customized search logs according to the user define input. This application will be processed in any of the search engine. In this paper, a new system called as Semantic Search log Social Personalized Search is also proposed which would be able to provide results for search query that relates to a particular user's environment, his area of interests, his likes and dislikes, the data the he/she might have found to be useful for him while searching. Social networks are the domain in which it could obtain such user oriented information, which could be used for providing personalized search results. Supervised learning technique is used to learn about the user based upon his interactions inside the system. This process can be able to make applicable for each and every registered user in this application. User can give their basic information in their profile and get benefits from their each and every search. When the user getting register with the system it creates an ontological profile, according to the profile created by the user and when he/she getting login into the social network and interacts with it the system updates his/her ontological profile based upon the interaction. The search provision can be finding out in their home page after they get login. When the user searches a keyword using the search engine inside the social network, it refers to the ontological profile of the user and displays the Personalized Search results. The system should be able to intelligently identify whether a search result has been useful to him or not and save it for his future reference when he searches for the same or similar keyword next time.***

Keywords: ontology search engine, semantic search, social personalized search, web crawler, Information Retrieval Process.

1. Introduction

A software program or script available through the Internet that searches document and files for keywords and returns the results of any files containing those keywords. There are thousands of different search engines available on the Internet, each with their own abilities and features. The first search engine ever developed is considered Archie, which was used to search for FTP files and the first text-based search engine is considered Veronica. The most popular and well known search engine is Google. The search engines not only just search the pages but also display the results depending upon their importance. This importance is commonly determined by using various algorithms.

Major search engines such as Google, Yahoo AltaVista, and Lycos index the content of a large portion of the Web and provide results that can run for pages - and consequently overwhelm the user. Specialized content search engines are selective about what part of the Web is crawled and indexed. For example, Tech Target sites for products such as the AS/400 (http://www.search400.com) and CRM applications (http://www.searchCRM.com) selectively index only the best sites about these products and provide a shorter but more focused list of results.

The Open Directory Project listed 370 search engines available for Internet users. There are about ten major search engines, each with its own anchor Web site (although some have an arrangement to use another site's search engine or license their own search engine for use by other Web sites). Some sites, such as Yahoo, search not only using their search engine but also give you the results from simultaneous searches of other search indexes. Sites that let you search multiple indexes simultaneously include:

Yahoo (http://www.yahoo.com)
search.com (http://search.com)
EasySearcher (http://www.easysearcher.com)

1.1 Yahoo

Yahoo first searches his own hierarchically structured subject directory and gives you those entries. Then, it provides a few entries from the AltaVista search engine. It also launches a concurrent search for entries matching your search argument with six or seven other major search engines. You can link to each of them from Yahoo (at the bottom of the search result page) to see what the results were from each of these search engines. A significant advantage of a Yahoo search is that if you locate an entry in Yahoo, it's likely to lead you to a Web site or entire categories of sites related to your search argument.

1.2 Search.com

A search.com search primarily searches the Info seek index first but also lets you search the other major search engines as well.

1.3 Easy searcher

Easy Searcher lets you choose from either the popular search engines or a very comprehensive list of specialized search engine/databases in a number of fields.

Yahoo, search.com, and Easy Searcher all provide help with entering your search phrase. Most Web portal sites offer a quickly-located search entry box that connects you to the major search engines.

2. Related Work

Innovation and agility should be provided to businesses by efficient collaboration (i.e., communication and sharing) between them. However, semantic heterogeneity between business processes is a serious problem for automatically supporting cooperation processes (e.g., knowledge sharing and querying-based interactions) between businesses. In order to overcome this problem, a novel framework is proposed based on aligning business ontologies for integrating heterogeneous business processes. Two types of alignment processes is considered; (i) manual alignment for building whole business process ontology in a business process management (BPM) system and (ii) automated alignment between business processes of different BPM systems. Thereby, the optimal integration between two business processes has to be discovered to maximize the summation of a set of partial similarities between semantic [4] components consisting of the business processes. In particular, the semantic components are extracted from semantic annotations of business processes. For evaluating the proposed system, we have conducted experimentations by using 22 business process management systems, which are organized as six business alliances. Here it is assumed that business processes in a same BPM system[9] should be built with common ontologies. The proposed alignment method has shown about 71.3% of precision (65.4% of recall). In addition, we found out that alignment results are dependent on some characteristics of ontology (e.g., depth and number of classes).

Semantics is applied into Business Process Management (BPM) to bridge the gap between the business world and information systems, especially in the context of B2B integration. Current standards such as BPMN, XPDL, BPEL and their combination have not fulfilled the expectation of two communities. The gap is still there: how enterprises can make the cross collaboration each other without 'knowing' their partners; and how a process based on the graphical notation can be fully mapped into the executable process without its semantics. In this paper, proposed a new approach, namely BizKB framework,[10] for the cross-enterprise integration using ontologies and Semantic Web Services technologies in order to realizing business concepts into the executable level using web services.

Business process management (BPM)[11][9] aims at supporting the whole life-cycle necessary to deploy and maintain business processes in organizations. Despite its success however, BPM suffers from a lack of automation that would support a smooth transition between the business world and the IT world. It is argue that semantic BPM, that is, the enhancement of BPM with semantic Web services technologies, provides further scalability to BPM by increasing the level of automation that can be achieved. The particular SBPM approach is developed within the SUPER project and illustrates how it contributes to enhancing existing BPM solutions in order to achieve more flexible, dynamic and manageable business processes.

Cross-enterprise collaboration [12] [13] is one of challenges on the business-to-business integration (B2Bi) research nowadays. With the support of Semantic Web technologies, the gap between business and IT communities has been reduced in order to tackle the mentioned challenge. Semantic Web-based approaches for BPM have been a promising solution with taking advantages of Semantic Web technologies such as ontologies, semantic web services. In this paper, a new approach is proposed called Ontological Hierarchical Task Network (O-HTN) based on HTN Planning and Web Service Modeling Ontology (WSMO) for forming collaborative business processes dynamically for the cross-enterprise collaboration.

This technical paper discusses the architectural design and implementation details of Genesis[12] - a novel Web application which formulates business process definitions dynamically, given a user business goal and underlying business criteria (e.g. total order cost, type of sourcing methods, etc.). Guided by an algorithm that references a hierarchical ontology file containing business process task decomposition, Genesis traverses through the ontology and dynamically produces two types of output: (a) a graphical breakdown of task sequences and decompositions required to fulfill the user's business goals, and (b) an abstract BPEL file containing the control flow structures, and Web service invocation points needed to execute the collaborative business processes in a service-oriented environment. The outputs demonstrate the potential of Genesis as a standalone module which can provide dynamic capabilities, thereby complementing current service-oriented architecture (SOA) business-to-business (B2B)[11] information systems which require hard-coded, inflexible business process definitions.

Increased trade and globalization has created an increasing need for the dynamic formulation and integration of cross-enterprise collaborative business processes (cBP's). However, current systems and methodologies, being static in nature, are unable to dynamically formulate cBP's based on business goals and selection criteria. Much of this stems from the current inability to bridge high level strategic business goals to low-level operational tasks, and the inability to dynamically decompose compound business process tasks into primitive operational tasks for direct Web service execution. This paper, demonstrate how the concepts from hierarchical task network (HTN) planning are feasible for dynamically creating cBP task sequences ideal for direct Web service execution. Also establish the rationale behind modeling business-to-business (B2B) collaboration tasks as hierarchical Web ontologies. To demonstrate the achievability of dynamic cBP formulation, we developed the genesis methodology, which consists of (a) business-OWL (BOWL) - a B2B hierarchical task Web ontology, and (b) the genesis algorithm - an extension of the hierarchical task network (HTN) planning algorithm to handle business criteria and control flows commonly found in business processes.

Cross-enterprise collaboration [10] is one of challenges on the business-to-business integration (B2Bi) research nowadays. With the support of Semantic Web technologies, the gap between business and IT communities has been reduced in order to tackle the mentioned challenge. Semantic Web-based approaches for BPM have been a

promising solution with taking advantages of Semantic Web technologies such as ontologies, semantic web services. In this paper, we propose a new approach called Ontological Hierarchical Task Network (O-HTN) [13] based on HTN Planning [12] and Web Service Modeling Ontology (WSMO) for forming collaborative business processes dynamically for the cross-enterprise collaboration.

Semantic business process management (SBPM) [9] emerges as a promising solution to the gap between businesses and information technology field with the approach to perform business actions which are supported by the information technology with perspective of business process rather than technical perspective. Managing business processes shall include methods, techniques and tools to support in designing and constructing rules as well as managing and analyzing businesses operations. However, handling the BPM automatically in integrating business processes among enterprises is still low due to the interaction between the business process collaboration's semantics.

To solve this problem, many researchers have recently proposed solutions in apply article intelligences in managing the processes of the collaboration between enterprises discussed. This paper proposes an approach called Ontological HTN (O-HTN) based on HTN Planning and Web Service Modeling Ontology (WSMO) for forming collaborative business processes [13] dynamically for the cross-enterprise collaboration.

3. Existing System

The existing system of this research is involved with the older search engine process. As well the existing system involves with a simple search engine using now a days. In a basic search engine, the web crawling is done by several distributed crawlers. There is a URL server that sends lists of URLs to be fetched to the crawlers. The web pages that are fetched are then sent to the store server. The store server then compresses and stores the web pages into a repository. Every web page has an associated ID number called a doc ID which is assigned whenever a new URL is parsed out of a web page. The indexing function is performed by the indexer and the sorter. The indexer performs a number of functions. It reads the repository; uncompressed the documents, and parses them. Each document is converted into a set of word occurrences called hits. The hits record the word, position in document, an approximation of font size, and capitalization. The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index. The indexer performs another important function. It parses out all the links in every web page and stores important information about them in an anchors file. This file contains enough information to determine where each link points from and to, and the text of the link.

The URL resolve reads the anchors file and converts relative URLs into absolute URLs and in turn into doc IDs. It puts the anchor text into the forward index, associated with the doc ID that the anchor points to. It also generates a database of links which are pairs of doc IDs. The links database is used to compute Page Ranks for all the documents. The sorter takes the barrels, which are sorted by doc ID, and resorts them by word ID to generate the inverted index. This is done in place so that little temporary space is needed for this operation. The sorter also produces a list of word IDs and offsets into the inverted index.

4. Information Retrieval Process

The main idea is to satisfy the user information need by searching over the available material for information that seems relevant. In order to accomplish this, the IR system consists on several modules that interact among them .It can be described, in a general form, as three main areas: Indexing, Searching, and Ranking.

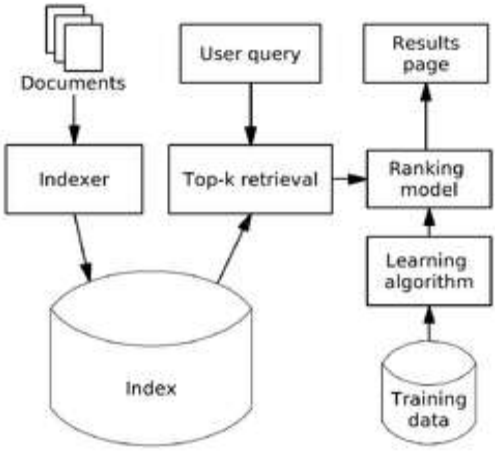

Figure 1: Information Retrieval process

4.1 Indexing

Various access methods have been developed to support efficient search and retrieval over text document collections. Inverted files have traditionally been the index structure of choice for the Web. Commercial search engines use custom network architectures and high performance hardware to achieve sub–second query response times using such inverted indexes.

4.2 Searching

In charge of extracting information from the index that satisfies the user information need.

4.3 Ranking

Although this is an optional task, it is also very important for the retrieval task. It is in charge of sorting the results, based on heuristics that try to determine which results satisfy better the user need.

5. Proposed System

According to the proposed system the web crawler will be concentrated more. This proposed system will be furnished in technical as follows:

Running a web crawler is a challenging task. There are tricky performance and reliability issues and even more importantly, there are social issues. Crawling is the most fragile application since it involves interacting with

hundreds of thousands of web servers and various name servers which are all beyond the control of the system.

In order to scale to hundreds of millions of web pages, Google has a fast distributed crawling system. A single URL server serves lists of URLs to a number of crawlers. Both the URL server and the crawlers are implemented in Python. Each crawler keeps roughly 300 connections open at once. This is necessary to retrieve web pages at a fast enough pace. At peak speeds, the system can crawl over 100 web pages per second using four crawlers. This amounts to roughly 600K per second of data. A major performance stress is DNS lookup. Each crawler maintains its own DNS cache so it does not need to do a DNS lookup before crawling each document. Each of the hundreds of connections can be in a number of different states: looking up DNS, connecting to host, sending request, and receiving response. These factors make the crawler a complex component of the system. It uses asynchronous IO to manage events, and a number of queues to move page fetches from state to state.

But this problem had not come up until we had downloaded tens of millions of pages. Because of the immense variation in web pages and servers, it is virtually impossible to test a crawler without running it on large part of the Internet. Invariably, there are hundreds of obscure problems which may only occur on one page out of the whole web and cause the crawler to crash, or worse, cause unpredictable or incorrect behavior. Systems which access large parts of the Internet need to be designed to be very robust and carefully tested. Since large complex systems such as crawlers will invariably cause problems, there needs to are significant resources devoted to reading the email and solving these problems as they come up.

5.1Fuzzy Methodologies

The main steps of the algorithm are:

- Sampling the graph.
- Bisecting the graph.
- Assigning the sampled documents to the best cluster found.
- Reapplying the algorithm on the two found partition.
- Choosing the best order of the two partitions.

For the analysis purposes let's assume that after the first sampling step we have a scaled down terms-documents bipartite graph Ge = (V, e Ee) with $|Ve| = N\rho,\tau$, where $N\rho,\tau$ is the number of documents remained after this initial step. Moreover let TMetis ($N\rho,\tau$) be the time spent by the Metis to compute a two-way splitting of the graph Ge. By simply summing the cost of each of the fuzzy algorithm we obtain the following recurrence formula for its complexity:

$T(N) = TMetis (N\rho,\tau) + 2 \cdot N\rho,\tau + 2 \cdot (N - N\rho,\tau) + 2 \cdot TN2=$
$= TMetis (N\rho,\tau) + 2 \cdot N\rho,\tau + 2 \cdot N - 2 \cdot N\rho,\tau + 2 \cdot TN2=$
$= TMetis (N\rho,\tau) + 2 \cdot N + 2 \cdot TN2>> 2 \cdot N + 2 \cdot TN2=$
$= N \cdot \log (N)$

For this reason the complexity of the fuzzy algorithm is $\Omega (N \cdot \log (N))$.

5.2 Generic top-down assignment algorithm

TDAssign(D, H e): the generic top-down assignment algorithm.

1: Input: The set De. The function H used to select the initial documents to form the centers of mass of the partitions.
2: Output: An ordered list representing an assignment function π for De.
3: De0, De00= HDe;
4: c1 = center of mass De0;
5: c2 = center of mass De00;
6: for all not previously selected d ∈ De \De0 ∪ De00do
7: if De0 ≥ |De|2∨De00 ≥|De|2then
8: Assign d to the smallest partition;
9: else
10: dist1 = distance (c1, d);
11: dist2 = distance (c2, d);
12: if dist1 < dist2 then
13: De0 = De0 ∪ {d};
14: else
15: De00 = De00 ∪ {d};
16: end if
17: end if
18: end for
19: Df0ord = TDAssign(Df0, H);
20: Df00ord = TDAssign(Df00, H);
21: if Df0ord Df00ord then
22: Deord = Df0ord ⊕ Df00ord
23: else
24: Deord = Df00ord ⊕ Df0ord
25: end if
26: return Deord

5.3 k-scan

The other bottom-up algorithm developed is k-scan. It resembles to the k-means one. It is, indeed, a simplified version requiring only k steps. At each step i, the algorithm selects a document among those not yet assigned and uses it as centered for the i-th cluster. Then, it chooses among the unassigned documents the $|De| k - 1$ ones most similar to the current centered and assign them to the i-th cluster. The time and space complexity is the same as the single-pass k-means one and produces sets of ordered sequences of documents. Such ordering is exploited to assign consecutive Doc IDs to consecutive documents belonging to the same sequence.

5.4 Crawling

Several papers investigate the design of effective crawlers for facing the information growth rate. There are several aspects aimed at improving classical sideling schemes. Main efforts are oriented toward finding an effective solution for reducing the number of sites visited by Crawlers. The main contributions are:

- URL–ordering
- Focused Crawling; and
- Incremental Crawling.

The URL–ordering technique consists of sorting the list of URLs to be visited using some importance metrics and in crawling the Web according to the established ordering. This

technique impact both the repository refresh time and the resulting index quality since the most important sites are chosen first. In, Garcia–Molina et al. investigate three importance measures to establish site importance: Similarity to a Driving Query Q, where the importance is measured as the distance among the URLs content and a query Q, Back link Count where the importance is the number of URLs linking to the current URL, and Page Rank which is based on the Page Rank ranking metrics. From the paper we could devise two main aspects.

In sideling algorithms which consider only Page Rank and Back link count, the Page Rank strategy outperforms the other due to its non–uniform traversing behavior: going in depth when the importance of the children is high enough, moving to the siblings whenever child nodes contain unimportant documents. On the other hand, when a similarity driven crawling algorithm is used the Page Rank strategy is comparable to the Breadth first traversal. This happens because when a page is authoritative with respect to a particular topic; its children are likely to have a high importance too. In their work the authors restricted the crawling space to the Stanford University Web pages.

Focused Crawling is an argument very close to URL Ordering. A focused crawler is designed to only gather a document on a specific topic, thus reducing downloads and the amount of network traffic. The crawler starts by using a canonical topic taxonomy and user specified starting points (e.g. bookmarks). A user marks interesting pages as he browses. Such links are then placed in a category in the taxonomy. This was bootstrapped by using the Yahoo hierarchy (260,000 documents).

The main components of the focused crawler are a classifier, a distiller and a crawler. The classifier makes relevance judgments on pages to decide on link expansion, and the distiller determines centrality of pages to determine visit priorities. The latter is based on connectivity analysis. In order to evaluate the proposal, authors consider the harvest ratio, i.e. the rate at which relevant pages are acquired, and how effectively irrelevant pages are filtered away. They state that is desirable to start from keyword-based and limited-radius search. Another observation was that the web graph is rapidly mixing: random links rapidly lead to random topics. At the same time, long paths and large sub graphs exist with topical coherence.

6. Experimental Results

The experimental results described in the paper show that the two–level caching generally outperforms the others. The two–level cache allows increasing the maximum throughput (the number of queries processed per second) by a factor of three, relative to an implementation with no cache. Furthermore, the throughput of the two–level cache is up to 53% higher than the implementations using just cache of inverted lists and up to 36% higher than the cache of query results.

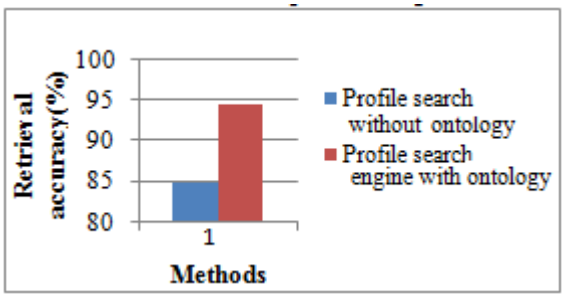

Figure 2: Retrieval accuracy comparison of profile search without and with ontology

7. Conclusion and Future Work

The design and implementation, as well the analysis, of efficient and effective Web Search Engines (WSEs), are becoming more and more important as the size of the Web has continually kept growing. Furthermore, the development of systems for Web Information Retrieval represents a very challenging task whose complexity imposes the knowledge of several concepts coming from many different areas: databases, parallel computing, artificial intelligence, statistics, etc.

Thus have implemented ontology concept [3] successfully as per committed in the introduction. Here the testing was done on the local host successfully. As per the future enhancement this concept can be implemented in web server or else in any server like cloud, grid and etc.

In future this research can be presented an analysis of several efficient algorithms for computing approximations of optimal assignment for a collection of textual documents that effectively enhances the compressibility of the index built over the reordered collection. The algorithms shown operate following two opposite strategies: a top-down approach and a clustering approach. In the first group fall the algorithms that recursively split the collection in a way that minimizes the distance of lexicographically closed documents.

The second group contains algorithms which compute an effective reordering employing linear space and time complexities. The experimental evaluation conducted with a real world test collection, resulted in improvements up to 23% in the compression rate achieved. The improved ontology can be presented by indexing with the proposal of a novel software architecture exploiting the parallelism among the phases of the indexing process.

References

[1] M. Petkovic and W. Jonker, “An Overview of Data Models and Query Languages for Content-Based Video Retrieval,”Proc. Int’l Conf. Advances in Infrastructure for E-Business, Science, and Education on the Internet,Aug. 2000.
[2] M. Petkovic and W. Jonker, “Content-Based Video Retrieval by Integrating Spatio-Temporal and Stochastic Recognition of Events,”Proc. IEEE Int’l Workshop Detection and Recognition of Events in Video,pp. 75-82, 2001

[3] Prote´ge´OntologyEditor,"http://protege.stanford.edu/, 2012.
[4] Jena:ASemanticWebFramework,"http://www.hpl.hp.com/semweb/, 2012
[5] U. Akdemir, P.K. Turaga, and R. Chellappa, "An Ontology Based Approach for Activity Recognition from Video,"Proc. ACM Int'l Conf. Multimedia,A. El-Saddik, S. Vuong, C. Griwodz, A.D. Bimbo, K.S. Candan, and A. Jaimes, eds., pp. 709712,http://dblp.unitrier.de/db/conf/mm/mm2008.html#AkdemirTC08, 2008.
[6] Y. Yildirim, "Automatic Semantic Content Extraction in Video Using a Spatio-Temporal Ontology Model," PhD dissertation, Computer Eng. Dept., METU, Turkey, 2009.
[7] Y. Yildirim and A. Yazici, "Ontology-Supported Video Modeling and Retrieval," Proc. Fourth Int'l Conf. Adaptive Multimedia Retrieval: User, Context, and Feedback (AMR),pp. 28-41, 2006.
[8] Y. Yildirim, T. Yilmaz, and A. Yazici, "Ontology-Supported Object and Event Extraction with a Genetic Algorithms Approach for Object Classification,"Proc. Sixth ACM Int'l Conf. Image and Video Retrieval (CIVR '07),pp. 202-209, 2007.
[9] Jason J. Jung "Semantic business process integration based on ontology alignment" Department of Computer Engineering, Yeungnam University, Dae-Dong, Gyeungsan 712-749, Republic of Korea.
[10] HanhHuu Hoang, "BizKB: A Conceptual Framework for Dynamic Cross-Enterprise Collaboration" ThanhManh LeComputational Collective Intelligence. Semantic Web, Social Networks and Multiagent SystemsLecture Notes in Computer Science Volume 5796, 2009, pp 401-412.
[11] Pedrinaci, C."**"**Semantic Business Process Management: Scaling Up the Management of Business Processes" Knowledge Media Inst., Open Univ., Milton Keynes .
[12] VuMinhHoang ThuaThien Hue Branch, Vietnam Posts &Telecommun., Hue, Vietnam HanhHuu Hoang "An Ontological Approach for Dynamic Cross-Enterprise Collaboration"Advanced Information Networking and Applications Workshops (WAINA), 2012.
[13] R.K.L. Ko, A. Jusuf, S.G. Lee"Genesis – Dynamic collaborative business process formulation based on business goals and criteria" Sch. of Mech. &Aerosp. Eng., Nanyang Technol. Univ., Singapore, Singapore.
[14] SomboonHongeng, Ram Nevatia and Francois Bremond "Video-Based Event Recognition: Activity Representation and Probabilistic Recognition Methods" University of Southern California" Institute for Robotics and Intelligent Systems Los Angeles, California 90089.