WEB AND THE INFORMATION RETRIEVAL: AHEAD OF CURRENT SEARCH ENGINES

Mohd Javed Mohd Nadeem Mansuri*

Prof R.R. Keole**

Abstract

In this paper we describe the challenges which are faced in information retrieval (IR) on the Web, in particular other types of data, what is Web mining and the issues relating to crawling. We also describe how IR and Soft Computing are related to each other and how these techniques describe the challenges.

* ME 1st Year (Computer Science and Engineering), H.V.P.M COET, SGB Amravati University

** Prof Information Technology, H.V.P.M COET, Amravati

1. Introduction

Nowadays the Web can be stated as the largest and easy available repository of data. Hence, it becomes obvious to extract information from it and Web search engines prove to be one of the most used tools in Internet. However, the ever increasing demand and that to at a very fast rate, makes really hard to retrieve all useful and required information. In fact, the main hindrance for Web search engines is crawling.

Recently done study on the challenges of searching the Web include the following Problems [22, 17, 8]:

- Exploiting user feedback, either from explicit user evaluation or implicitly from Web logs. We can include here implicit information given by the authors of Web pages in the form of several conventions used in HTML design.
- Improving the query language, adding the context of the information needed, such as genre or time.
- Identifying content of good quality. The Web mostly includes low quality (syntactic and semantically) content, including noisy, unreliable and contradictory data. Hence, we have the problem of how much a Web site can be trusted. This includes HTML structure, which in most cases is vague and heterogeneous.
- Detecting duplicate hosts and content, to avoid unnecessary crawling.
- Identifying and removing malicious content and linking, called search engine spam. Some authors call this problem adversarial IR.
- Keeping the index fresh and complete, including hidden content.
- Improving ranking, in particular to make it dependent on the person posing the query. Relevance is based in personal judgments, so ranking based in user profiles or other user based context information can help. Here we
  Can add quality, trust, and user feedback issues.

All these problems can be better understood when we search for real data. Thus we can add much more results about it. Additional material can be found in [12, 14, 13].

The Web is more than plain HTML and other text dominant formats and we would like to search well other data types. Among them we have dynamic pages, multimedia objects, and XML data and associated semantic information. If

The Semantic Web becomes a reality in spite of all the social issues that need to be solved, we may have an XML-based Web, with standard semantic metadata and schema. In that possible world, information retrieval (IR) becomes easier, And even multimedia search is simplified. Spam should disappear in this setting and it is easier to recognize good content. On the other hand, new retrieval problems appear, such as XML processing and retrieval, and Web mining on Structured data.

The idea of soft computing (SC) was initiated in 1981 by Lofti A. Zadeh [9] as an aim to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve close resemblance with human like decision making. Zadeh defined Soft Computing into one multidisciplinary system as the fusion of the fields of Fuzzy Logic, Neuro-Computing, Evolutionary and Genetic Computing, and Probabilistic Computing. The Main characteristics of Soft Computing is that Soft computing systems use Fuzzy logic, which offers a flexible approach to dealing with the human-like classification of things into groups whose boundaries are vague and also provide approximate reasoning and explanations. Another advantage is that artificial neural network offers ability to learn. The vagueness, and imprecision are typical properties of any IR process. Hence SC techniques are acceptably used to improve IR process. It can be better stated that, it’s an application to solve the different IR problems recently appeared in the Web can be of help.

We start by having a short description to Web Mining, followed by the explanation of Data Challenges, and then continuing with some ways that can help partially to solve the crawling problems. We end with a short description on the use of SC in IR.

2. Web mining:

Data mining is the search for relationships and patterns that exist in large databases, but are 'hidden' among the vast amounts of data and we don’t know how to retrieve it. But in IR we know the query to retrieve it. Hence, we try to find relations in the data that look like an interesting answer and then we study it to find the corresponding query. In the Web this leads to Web mining, a further challenge beyond IR on the Web. Some authors misjudge and describe IR as Web Mining. But we believe it to be incorrect. Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents web sites, etc.

We can observe that Web mining completes three main tasks: structure, usage and content. Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site. Content is the process of extracting useful information from the contents of Web documents. Structure is the process of discovering structure information from the Web. Also not only the three cases stated before are present but also we have a temporal dimension defining how the dynamics of the Web growths and changes. This implies temporal data. The first two types are covered in [5], while the third is the main topic of [14]. The later type is less studied and some results are presented in [2].

In addition to find new information or knowledge Web mining can be used for several purposes. It can be used for adaptive Web design (for example, user-driven Web design), Reorganization of web site, Web site personalization, and for making various performance improvements. However, the search engine face the same problem today also: collecting the data.

3. Improving the web search engine:

For a search engine to be perfect it should solve the problems mentioned before, could retrieve any type of data and should collect information in order to do a better web mining. The problem of crawling consist of the volume of data present on web and the growth of that data, these issues together teaming up with duplicated and volatile data, and a very inefficient technique called pulling.

Current Search Engine sends out programs, which are known in the trade as "robots" (because they work automatically) or "spiders" (because they crawl the World Wide Web.) These programs gather all the information on your website that they can find, or that you allow them to see but without having any interaction with the Web Servers. They must transfer the pages using the standard HTTP protocol through TCP ASCII connections, and poll them to see if a page has been modified, to update their indexes after pulling updated or new pages.

It is more useful to send an agent to the server, where new pages, links, and modified pages can be searched at local level. It can also pack together all updated pages in a compressed file so that they can be transferred to the search engine. The main search server can guide the remote agent so as to help him decide whether it is worth to transfer the existing batch based on several parameters such as the number of files, importance of them, etc. The crawler can then be useful by providing equal intelligence to the main search engine and the existent agents. [23] Study the impact on the bandwidth when Web servers publish metadata of their Web pages, such as actualization dates, size, etc. They show that

There are savings and also the freshness of the pages increases. A similar paper focuses on freshness [24]. But we can make a better improvement by going one step further and pushing information rather than pulling it.

The topic of interaction then shifts from pulling pages to pushing changes. But the other extreme is also not efficient, as pushing too much will overload the main central server. So, the best option is to let the server do all the negotiation prior with the agent as to when to send a message and how to send a message giving an alarm that a batch of changes is ready (or even better, that the changes have been already indexed and a partial index is available). The main server will then pull at due time those changes. This creates a long-term scheduling, to which more changes can be made when the Web server that pushed a warning is visited. Instead of all, this scheduling still proves to be simpler than the current ones as we have more information, and we don’t have to worry about politeness as we can have surety that all accesses are not frequent and they are always successful.

In general, Web servers will want to cooperate in this architecture, because today it is an accepted value to be indexed on a popular search engine. On the other hand, even if the CPU cycles are spent on behalf of the search engine, the crawler is not being polling them, thus Web server access load gets effectively diminished. Also when the load is lower these cycles can be spent in that periods.

As a first testing stage, when there is no availability of a global available agent platform, a simple module, associated to the Web server, which provides a similar functionality and measures the performance improvement could be developed. As we have already stated, suggestion of small changes to the Web server that can enable cooperation with search engines

[23, 24] have already been done, but they lack flexibility and they interfere with the crawler policies. Agents can provide a good help to improve such type of behaviour, so that their algorithms can get priories pages in order to get embedded in the agents code. In this sense, the agent proves to be an important component in the crawler's algorithm, and a particular search engine's policies are followed by its logic [4].

4. Challenges in Data:

There is a need to address several data issues. Among them we have to mention multimedia data, semantic data, structured data, and hidden or dynamic pages. After that we describe each one of them, except hidden data, that is a particular case of generic data with the problem of restricted access.

4.1. Multimedia Data:

Multimedia data includes images, animations, audio in several forms, and video. All of them have no standard formats. Considering the images the dominant ones are JPG, GIF and PNG, for music the most popular is MP3, MP4 Real Video or QuickTime for video, etc. The ideal solution would be to search on any kind of data, which include text, by using same model and with one single query language. This ambitious goal is probably not possible.

A similar model can be developed for a particular data type, and the query language will change accordingly depending on its type. For instance, query by humming for audio query by example for images. Rather than belonging to classical IR, all this area belongs more to image and signal processing.

4.2. Semantic Data:

The two main problems with semantic information are reduced anonymity on the Web and increased invasion of privacy. The first is being carried out by the WWW Consortium while the second one requires schemes for certification which could be developed in future.

Other problems are common issues such as intelligent content scraping, distributed authority, heterogeneous content and quality. An introduction to these and other challenges for the Semantic Web are presented in [20, 25, and 18].

4.3. Structured Data:

Data that resides in a fixed field within a record or file is called structured data. Examples are e-mail, news postings, etc. If XML becomes prevalent, the structure level is even higher. The first challenge is to design data models and

Associated query languages that allow to mix content and structure. Structured text was considered before XML and several efficiency/expressiveness trade-offs were designed [3]. After XML, the WWW Consortium has proposed XQuery as standard [28].

When retrieving XML data several challenges are to be faced:

- A single XML object can contain many answers, and there is possibility of overlapping.
- The answer can be a fragment of XML and not necessarily a complete object. Nevertheless, the answers should also be XML based data.
- How we can rank an answer and how ranking is inherited if we need to project the answer to certain structure types? Sometimes the combination of sub trees should have a better ranking if they are close, but in other cases is good if they are far apart.

Recent research on these topics is given in [19, 21, 11, and 6]. An additional problem is processing XML streams. That is, filtering a stream of XML objects with a large set of queries. Here the queries can be indexed, but not the data. See [7] for an introduction to this problem.

4.4. Dynamic Data:

As compared to content generated on demand the static Web proves to be smaller, in particular by querying in e-business or information services sites. Dynamic links are followed by Current crawling software, but care has to be taken while doing it, as the same page can be generated again and again or even it might go unlimited. Since the crawler does not have any knowledge of database accessing pages behind query forms becomes more difficult. On the other hand, even if the database is known, asking all possible queries might consume a lot of time (exponential on the size of the database) and even if we stick to simple queries, real persons would never posed some of them. Even if they allow to learn from the database and how people query in it, web services might prove be a partial solution to this problem. For example, obtaining the most frequent one thousand queries could be enough. Another possibility is analysing the page as in [26].

5. Information retrieval and Soft computing:

IR aims at obtaining information resources relevant to an information need from a collection of information resources.

It helps to model, design, and implement system that can be able to provide fast and effective content-based access to large amount of information. The aim of an IR system is to estimate the relevance of information items to a user

Information need expressed in a query. Since it is pervaded with subjectivity, vagueness and imprecision, it is becomes a very hard and complex task.

As we have already mentioned above in the starting that the idea of soft computing (SC) was initiated by Lofti A. Zadeh [9]. The main aim of soft computing is to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve close resemblance with human like decision making. A set of techniques is provided by SC for appropriately handling vagueness, subjectivity, and imprecision which exist in several problems.

SC is defined into one multidisciplinary system as a fusion of fields of Fuzzy Logic, Neuro-Computing, Evolutionary and Genetic Computing, and Probabilistic Computing. The IR problem is a typical application field of SC. Some main approaches of IR in SC are given below:

- Probabilistic techniques: ranking, web mining.
- Fuzzy sets and logic: information fusion, text extraction, query language models, and document clustering.
- Bayesian networks: retrieval models, ranking, thesaurus construction, and relevance feedback.
- Genetic algorithms: document classification, image retrieval, relevance feedback, and query learning.
- Rough sets and multivalued logics: document clustering.

At least one hundred papers devoted to the problems just enumerated are made, and listing all of them is matter of a full survey. Nevertheless, we refer the reader to Miyamoto's book [23] as well as the excellent volume edited by

Crestani and Pasi [17], a special issue of IP&M [15], a survey paper by Chen [10] and this journal issue.

With the techniques mentioned above, at least half of the problems which are described in the introduction and in the subsequent sections can be addressed. Hence, there still remains a lot of research work which should to be done ahead. The main drawbacks could be answer explanation (that is, for example, why a document is classified in a given category?) and performance issues (for example, can be used in practical settings with bounded answer time?). Recent applications of SC to IR on the Web have been done that includes adaptive agents, user profiles, categorization of web age, quality assessment, etc. Hence, this shows that there is a lot of possibility that can help to progress in Web IR by the use of SC techniques.

6. Conclusion:

In this paper we discussed various types of challenges that are faced while retrieving information from web. We also discussed some of the possible method that can be used to make the information retrieval faster and easier. A brief discussion was done on how Information Retrieval and Soft Computing together define these challenges and also the web mining topic was described.

7. References

[1] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, England, 1999, 513p.

[2] R. Baeza-Yates, F. Saint-Jean, C. Castillo, Web Dynamics, Structure and Page Ranking, SPIRE 2002, Springer LNCS, Lisbon, Portugal, 2002, pp. 117–130.

[3] R. Baeza-Yates, G. Navarro, Integrating contents and structure in text retrieval, SIGMOD Record 25 (1) (1996) 67–79.

[4] R. Baeza-Yates, J. Piquer, Agents, Crawlers, and Web Retrieval, CIA 2002, Springer LNIA, Madrid, Spain, 2002, pp. 1–9.

[5] R. Cooley, B. Mobasher, J. Srivastava, Web mining: information and pattern discovery on the world wide web, ICTAI, 1997, pp. 558–567.

[6] L. Fegaras, R. Elmasri, Query engines for web-accessible XML data, The VLDB Journal (2001) 251–260.

[7] D. Suciu, from searching text to querying XML streams, SPIRE 2002, Lisbon, Portugal, pp. 11–26.

[8] H. Tirri, Search in vain: challenges for internet search, IEEE Computer (January) (2003) 115–116.

[9] L.A. Zadeh, Fuzzy logic, neural networks and soft computing, Communication of ACM 37 (3) (1994) 77–84.

[10]H. Chen, Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms, Journal of the American Society for Information Science 46 (3) (1995) 194–216.

[11]P. Fankhauser, A. Halevy, XML data management, The VLDB Journal 12 (1) (2003) (Special issue).

[12]A. Arasu, J. Cho, H. Garcia-Molina, S. Raghavan, Searching the web , ACM Transactions on Internet Technologies 1 (1) (2001).

[13]S. Chakrabarti, Recent results in automatic web resource discovery, ACM Computing Surveys (1999).

[14]S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data, Morgan Kaufmann, 2003.

[15]F. Crestani, G. Pasi (Eds.), Handling vagueness, subjectivity and imprecision in information access, Information Processing and Management, 39(2) (2003).

[16]F. Crestani, G. Pasi (Eds.), Soft Computing in Information Retrieval: Techniques and Applications, Physica-Verlag, Heidelberg, 2000.

[17]M. Henzinger, R. Motwani, C. Silverstein, Challenges in web search engines, SIGIR Forum 36 (2) (2002).

[18]F. van Harmelen, How the Semantic Web will change KR: challenges and opportunities for a new research agenda, The Knowledge Engineering Review 17 (1) (2002).

[19]R. Baeza-Yates, D. Carmel, Y. Maarek, A. Sofer (Eds.), IR and XML, JASIST 53(6) (2002) (Special issue).

[20]R. Benjamins, J. Contreras, O. Corcho, A. Gomez-Perez, Six Challenges for the Semantic Web, KR2002 Workshop on Formal Ontology, Knowledge Representation and Intelligent Systems for the Web, Toulouse, France, 2002.

[21]R. Baeza-Yates, N. Fuhr, Y. Maarek, Organizers. SIGIR Workshop on XML and IR, Tampere, Finland, 2002.

[22]O. Brandman, J. Cho, H. Garcia-Molina, N. Shivakumar, Crawler-friendly web servers, in: Workshop on Performance and Architecture of Web Servers (PAWS), June 2000.

[23]O. Brandman, J. Cho, H. Garcia-Molina, N. Shivakumar, Crawler-friendly web servers, in: Workshop on Performance and Architecture of Web Servers (PAWS), June 2000.

[24]V. Gupta, R. Campbell, Internet search engine freshness by web server help, Technical Report UIUCDCS-R-2000-2153, Digital Computer Laboratory, University of Illinois at Urbana- Champaign, January 2000.

[25]S. Lu, M. Dong, Ming, F. Fotouhi, The semantic web: opportunities and challenges for next generation web applications, Information Research 7 (4) (2002).

[26]S. Raghavan, H. Garcia-Molina, Crawling the hidden web, in: 27th International Conference on

Very Large Data Bases, September 2001.

[27]S. Miyamoto, Fuzzy Sets in Information Retrieval and Cluster Analysis, Kluwer Academic Publishers, 1990.

[28] WWW Consortium. XML Query Language Proposal. Available from <http://www.w3.org/XML/Query/>.