Advances in Focused Retrieval: A General Review

Shafiq Rayhan Joty [*], Sheikh Sadid-Al-Hasan [†]

[*] Department of Computer Science, University of Lethbridge, Lethbridge, Alberta, Canada, e-mail: *jotys@cs.uleth.ca*

[†] Department of Computer Science & Engineering, BUET, Dhaka, Bangladesh, e-mail: *sadid_hasan@yahoo.com*

***Abstract*—When a user is served with a ranked list of relevant documents by the standard Document Retrieval systems (i.e. search engines), his search task is usually not over. The next step for him is to look into the documents themselves in search for the precise piece of information he was looking for. This method is time consuming, and a correct answer could easily be missed, by either an incorrect query resulting in missing documents or by careless reading. Focused Retrieval tries to remove the onus on the end-user, by providing more direct access to relevant information. Focused retrieval is becoming increasingly important in all areas of information retrieval. In this paper we investigate the various aspects of Focused Retrieval.**

***Keywords*—Focused Retrieval, Information Retrieval, Passage Retrieval, Question Answering, Element Retrieval, Keywordese.**

I. INTRODUCTION

The size of the publicly indexable world-wide-web has provably surpassed several billions of documents and as yet growth shows no sign of leveling off. Dynamic content on the web is also growing as time-sensitive materials, such as news, financial data, entertainment and schedules become widely disseminated via the web. Search engines are therefore increasingly challenged when trying to maintain current indices using exhaustive crawling. Even using state of the art systems such as AltaVista's Scooter, which reportedly crawls ten million pages per day, an exhaustive crawl of the web can take weeks [7]. This vast raise in the amount of online text available and the demand for access to different types of information have, however, led to a renewed interest in a broad range of Information Retrieval (IR) related areas that go beyond simple document retrieval, such as focused retrieval, topic detection and tracking, summarization, multimedia retrieval (e.g., image, video and music), software engineering, chemical and biological informatics, text structuring, text mining, and genomics [5]. Focused Retrieval (FR) is relatively a new area of research which deals with retrieving specific information (i.e. Passage or Answer to a question or XML element) to the query rather than state of the art information retrieval systems (search engines), which retrieve documents. This means that the Focused Retrieval systems will possibly be the next generation of search engines. What is left to be done to allow the Focused Retrieval systems to be the next generation of search engines? The answer is higher accuracy and efficient extraction.

In this paper, we investigate various aspects of the three Focused Retrieval applications: a) Question Answering b) Passage Retrieval and c) Element Retrieval. The next section reviews theory. Section 3 provides the applications of the focused retrieval, Section 4 describes the interaction issues, section 5 describes the experience, and Section 6 concludes the paper.

II. THEORY

The increasing interest in sophisticated information retrieval (IR) techniques has led to a number of large collections becoming available for research. The size of these collections (both in terms of number of documents in them, and the length of the documents that are typically full text) and the format has presented significant challenges to IR researchers who are used to experimenting with two or three thousand document abstracts [1].

Our motivation is found in the integration of heterogeneous data. In order to carry out research with types of text representations, retrieval models, learning techniques, and interfaces, a new generation of powerful, flexible, and efficient retrieval engines needs to be implemented. Research in Data Integration [6] is being done to extract properties from unstructured texts and translate information into a common object model, that combine information from several sources, allow browsing of information, and that manage constraints across heterogeneous sites. Markup is used to represent different levels of granularity of text objects and to facilitate the retrieval. From the perspective of the search engine, there is no difference at all between a user query that contains keywords and stop words and a user query that just contains the keywords. As a consequence, as users gain experience with these search engines, they learn that the stop words don't have any value, and they just save themselves the trouble of typing them in their queries. The result is that the effective query language used by users trained in free-text queries is not natural language, but rather a keyword sequence language, which is called "keywordese". Most novice searchers, and even skilled searchers who are frustrated in a search session, still put stopwords in their queries, but skilled searchers consider this to be just silly.

But is it really silly to want to use stop words? On the contrary, we think this perspective reveals something fundamental about the state of search today. To begin with, let's look at the words that are "stopwords", the words that don't get any respect by the search engines. They are the little function words that put together the meaning of a phrase in a natural language like English. They are little because they are so frequent and useful in language. Words like "by", "for", "about", "of", and "in"

are all stopwords. But consider how valuable they are to communicating intent among humans. To a keywordese search engine, "book for programmers", "book by programmers", and "book about programmers" are all equivalent to "book programmers".

This motivates the idea of true natural language search. Instead of keywordese or even advanced keywordese, true natural language queries have linguistic structure. This includes queries where the function words matter, where word order means something, and where relationships that should be explicitly stated easily are stated. Instead of ignoring the function words, a natural language search engine respects their meaning and uses it to give better results. The notion of "Focused Retrieval" can be used as a label for a wide range of applications [21]:

Passage Retrieval is the task of identifying and extracting most relevant fragments from heterogeneous full-text documents.

Element Retrieval is a special case of passage retrieval where the passages are defined in terms of document structure (XML-IR). Element retrieval is used to search XML documents and identify relevant XML elements.

A **Question Answering** engine can be considered as "focused" since it finds answers to the questions. Historically different QA methodologies have been tried to achieve better accuracy while keeping the interaction efficient.

So, These Focused Retrieval systems raise interesting challenges to the NLP community. The systems can be benefited from the structural retrieval. NLP is trying to create richly annotated corpus and that will undoubtedly pose new challenges to XML retrieval and query languages.

III. APPLICATION

Due to the basic problems in Querying Mechanism, Keyword Exact Matching, Low web Coverage Rate, Long result list with low relevancy to user query and also because of the need to include information on specific domain, domain-specific search engines incorporating Focused Retrieval (Question Answering, Passage Retrieval, Element Retrieval) techniques are proposed.

While being quite successful in providing keyword based access to web pages, commercial search portals still lack the ability to answer questions expressed in a natural language [20]. Recent studies [18] indicated that the current WWW search engines offer a very promising source for open domain question answering. There are several question-answering sites on the Web: Two that come to mind are : 1) Brainboost Answer Engine[1] , which taps the answer.com reference site and answers questions related to geography, history, nutrition, people, science, computers, business, entertainment and product. 2) START[2] , MIT Professor Boris Katz's site that has been up since 1993. The first commercial company offering QA services, AskJeeves, supports natural language queries and provides the technological support for Ford and Nike [14]. A recent work by [9] presented an open-domain Web QA system that applies simple combinatorial permutations of words (so called "re-writes") to the snippets returned by Google and a set of 15 handcrafted semantic filters to achieve a striking accuracy: Mean Reciprocal Rank (MRR) of 0.507. The Initiative for the Evaluation of XML Retrieval (INEX) provides the infrastructure for conducting (XML) element retrieval experiments [10, 11]. So far, there has been no consensus about what a real-world application of element retrieval might look like, which was identified as a major obstacle to realistic experiments in this area [26]. Two commercial online digital library systems that provide search functions similar to those used at the INEX workshops offer full-text search for their online books, Books24x71 (launched in 1999) and Safari2 (launched in 2001)[8]. Much of all these active researches in focused retrieval are carried out by abstracting the user away from the problem: judgments are captured and held constant, non-binary relevance is ignored as too complex, evaluation is on a single round of query-results without opportunity to adjust the query, and so on. This abstraction has been incredibly successful in enabling research to advance rapidly, creating more effective and efficient systems for retrieving and organizing information. However, those improvements in retrieval accuracy appear to have dwindled in the past half dozen years. It may be that one reason researchers are unable to advance beyond the current plateau is that the evaluation model forces systems toward user-generic approaches that are "good enough" for everyone, and therefore "never great" for anyone [5]. We claim that greater focus on the user will enable major advances in focused retrieval technologies, perhaps dwarfing those made possible by better core algorithms. So, the bottom line is user modeling of focused retrieval applications will definitely extract un-addressed problems besides formally characterizing the existing difficulties in this arena.

The main motivation for the usefulness of focused retrieval system is based on the presence of the context scenario of nested information that fulfills the user's need appears locally nested within a longer document in the collection i.e., full documents are too long to be considered as the appropriate units of retrieval. The relevance of the focused retrieval approach depends on a number of "context variables," such as the data being searched, the person searching, and the task underlying the search [21].

All text structure comes in different kinds such as linguistic structure, document structure, and layout structure. Some structure is implicit, such as a chain of arguments that the author uses to tell her story. The other one is explicit text structure, such as paragraph segmentation, or assigned metadata. The markup of text documents can be used to represent different levels of granularity of text objects. Document structure can be marked-up using a number of different markup formats, such as, Microsoft-Word format [MS-Word], Portable Document Format [PDF], a scientific document preparation style [LATEX], HyperText Markup Language (HTML) [19], etc. Another general markup language, namely the eXtensible Markup Language (XML) [4] is a flexible markup language which serves as a representative example of modern semi-structured markup languages. Several different measures have been proposed to quantitatively measure the performance of classical information retrieval systems, most of which can be straightforwardly extended to evaluate focused retrieval methodologies as well. The system performance can be measured and compared using two types of information

[1] http://brainboost.com/
[2] http://start.csail.mit.edu/

retrieval evaluation frameworks, laboratory tests and operational tests [25]. A laboratory test is one where many environmental variables are controlled, while an operational test is one where none is controlled. Laboratory tests are useful for comparing systems or individual aspects of a single system—and are thus often referred to as systems-oriented evaluation.

A basic model from traditional retrieval systems recognizes a three-way trade-off between the speed of information retrieval, precision and recall. In the context of information retrieval [12], Precision measures the number of relevant documents retrieved as a portion of the total number of documents retrieved and Recall measures the number of relevant documents retrieved as a portion of the total number of relevant documents and several others used combining these two are F-measure, Precision-Recall curve, Average Precision (AP), Precision@N, R-Precision, Bpref etc [21]. So the baseline of measuring performance of the focused retrieval systems can be speed and accuracy parameters, where 'speed' will mean how fast response time the system offers and 'accuracy' will be determined based on relevance aspect of the retrieved information.

IV. INTERACTION

The effectiveness of end user interfaces depends solely on the fact how the design is focused to meet user demands in a quick response time. The main goal here is to have an improved environment than traditional document retrieval methods by gaining better understanding of the way in which users interact with a focused retrieval system. The end user interface demonstrates practicality as the system logs various interactions between the user and the system [22] (i.e. queries posted by users, information about which links on the result pages are clicked on by the user, all internal navigation pages etc). This data can be used to better understand how users interact with the system.

The laboratory evaluation framework has been criticized for its failure to account for user interaction. Interactive information retrieval evaluation studies the interaction between searchers and retrieval systems and compliments the laboratory evaluation framework [2, 3]. Interactive retrieval has been a part of TREC from its early days. The interactive track has developed four focal points [16]: 1. The searcher in interaction with the system, 2. Behavioral details, the process, and interim results not just summary measures of final result, 3. Isolation of the effects of system, topic, searcher, and their interactions and 4. Evaluation of the evaluation methodology. In the first TREC years, interactive systems were compared against automatic systems, but later the focus changed to comparing interactive systems among themselves [21]. The track collects data both on user satisfaction and the search process, including video, think-aloud audio, and system interaction logs. In TREC 1–8 the data was collected by assigning subjects a description of an information need and asking them to find as many relevant documents as possible within a given time period. The interactive track did not address any central research questions, but served as an experimental framework where participants could address their own questions. The participants did however share tasks, topics, documents, and assessments.

The results of [15] and [23] show that different user groups utilize search tools in different ways. Since the focused retrieval is a departure from the "traditional" document retrieval scenario, in the very beginning experienced searchers are targeted. Experienced searchers are more likely to be able to understand and utilize the new—and presumably more powerful—search approach. Focused information retrieval system is most likely to be useful for informational tasks. It is undisputed that in the research domain users need a quicker and more effective way to evaluate the content of retrieved documents [24].

A focused retrieval system should offer a structured overview of individual search results to be potentially useful than those offered in the standard text retrieval methods. To this end, a focused system could be useful by giving an overview of the content of different sub-parts of the documents. The results of [13] stress the importance of showing relevant information in context. A focused system should not take the sub-document-level retrieval results out of their document context, which is one of the limitations of document retrievals. A focused retrieval system should extend the state-of-the-art text retrieval approach with two new features [21]:

Structured result lists: A clear indication about the relation between the user's query and the "discourse structure" of the document is given. i.e., instead of showing only one text snippet for each document a text snippet for each relevant element is to be shown, together with a partial "table of contents".

Direct linking: Direct access is given to relevant portions i.e. relevant elements of documents. Following these links, users can get to the relevant information with less effort and less time.

V. EXPERIENCE

Information retrieval test collections provide a means to compare the effectiveness of different retrieval strategies in a laboratory setting. The most common test collections are based on the concept behind the Cranfield experiments [21]. The Cranfield paradigm has been applied in many settings, such as, TREC[3] , CLEF[4] , and INEX[5] .

Ad-hoc information retrieval collections usually consist of three parts: i) **Documents:** A collection of documents, over which search is performed, ii) **Topics:** Description of an information need. The information need is expressed in different formats—ranging from a short list of keywords to a verbose narrative, and iii) **Assessments:** A mapping between topics and documents indicating which documents satisfy the information need in the topic.

The Text REtrieval Conference (TREC) is an on-going series of workshops focusing on a list of different information retrieval tracks. Its purpose is to support and encourage research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies and to increase the speed of lab-to-product transfer of technology. The tracks of TREC serve several purposes. First, tracks act as incubators for new research areas: the first running of a track often defines what the problem really is, and a track creates the necessary infrastructure

[3] http://trec.nist.gov/
[4] http://www.clef-campaign.org/
[5] http://inex.is.informatik.uni-duisburg.de/

(test collections, evaluation methodology, etc.) to support research on its task. The tracks also demonstrate the robustness of core retrieval technology in that the same techniques are frequently appropriate for a variety of tasks. Finally, the tracks make TREC attractive to a broader community by providing tasks that match the research interests of more groups. Depending on track, test problems might be questions, topics, or target extractable features. Uniform scoring is performed so the systems can be fairly evaluated. After evaluation of the results, a workshop provides a place for participants to collect together thoughts and ideas and present current and future research work. TREC claims that within the first six years of the workshops, the effectiveness of retrieval systems approximately doubled. We have already mentioned that Interactive retrieval has been a part of TREC from its early days and Since 1999 TREC has the QA track with the goal to achieve more information retrieval than just document retrieval by answering factoid, list and definition-style questions.

CLEF stands for Cross-Language Evaluation Forum and it supports global digital library applications by developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts. The increasing use of XML, especially in scientific data repositories, Digital Libraries and on the Web, brought about an explosion in the development of XML systems, and in particular systems to store and access XML content. Whereas many of today's systems still treat documents as single large (text) blocks, XML offers the opportunity to exploit the logical structure of documents, which is explicitly represented by the XML markup, in order to allow for more precise access by giving more specific answers. Evaluating the effectiveness of XML retrieval systems, hence, requires a test collection where the relevance assessments are provided according to a relevance criterion, which takes into account the imposed structural aspects. A test collection as such has been built as a result of three rounds of the Initiative for the Evaluation of XML Retrieval (INEX 2002, INEX 2003 and INEX 2004). This initiative provides an opportunity for participants to evaluate their XML retrieval methods using uniform scoring procedures and a forum for participating organizations to compare their results. As part of a large-scale effort to improve the efficiency of research in information retrieval and digital libraries, this project initiated an international, coordinated effort to promote evaluation procedures for content-oriented XML-retrieval. In INEX 2005 and 2006 participating organizations were able to compare the retrieval effectiveness of their XML document retrieval systems and contribute to the continuous construction of a large XML test collection. The test collection will also provide participants a means for future comparative and quantitative experiments.

Interactive XML retrieval has been studied at INEX since 2004[21]. The main aim of the task has been to study the behavior of searchers when interacting with components of XML documents. Several dimensions determine complexity and difficulty to the QA systems: i) **Questions:** closed domain vs. open-domain. ii) **Data:** structured (e.g., relational) vs. semi-structured (XML) vs. unstructured (e.g., flat text). iii) **Answers:** extracted (e.g., text snippets) vs. generated (e.g., dialog). The process of examining what XML can and cannot do for QA will in turn point out challenges and issues for XML retrieval: several challenges for XML systems [17]:

1. Queries represent information needs and results should be ranked according to relevance.
2. There is a need to represent term weighting in XML queries.
3. Order and proximity are important for QA and this should be reflected in queries and indexing.
4. There are multiple overlapping hierarchies of structure for text that we would like to index.
5. It is important to be able to express relationships between components in these hierarchies.
6. The relationships between hierarchies may not be exact due to errors in language processing so approximate matching on structure and terms is important and should be reflected in the result rankings.

VI. CONCLUSION

In this discussion paper we tried to answer various questions on Focused Retrieval including its application, interaction and experience related issues. We believe that this discussion will be fruitful to the IR researchers who have been working in different areas of Focused Retrieval.

REFERENCES

[1] Abiteboul. S. “Querying semi-structured data” In Proceedings of ICDT, Jan 1997.

[2] Beaulieu M., Robertson S. E. and Rasmussen E. M. “Evaluating interactive systems in TREC”, Journal of the American Society for Information Science, 47(1):85–94, 1996.

[3] Borlund. P. “The IIR evaluation model: a framework for evaluation of interactive information retrieval systems”, Information Research, 8(3), 2003.

[4] Bray. T., Paoli. J. and Sperberg-McQueen. C. M., editors. “Extensible Markup Language (XML) 1.0.” W3C, 1998. http://www.w3.org/TR/1998/ REC-xml-19980210.

[5] "Challenges in Information Retrieval and Language Modeling", Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002.

[6] Chawathe. S., Garcia-Molina. H., Hammer. J., Ireland. K., Papakonstantinou, Y., Ullman. J. and Widom. J. “The TSIMMIS Project: Integration of Heterogeneous Information Sources”. In Proceedings of 10th Anniversary Meeting of the Information Processing Society of Japan, pages 7--18, Tokyo, Japan, 1994.

[7] Diligenti. M., Coetzee. F. M., Lawrence. S., Giles. C. L. and M. Gori, “Focused Crawling Using Context Graphs”, NEC Research Institute, 4 Independence Way, Princeton, NJ 08540-6634 USA.

[8] Dopichaj. P. “Element Retrieval in Digital Libraries: Reality Check”. In Proceedings of SIGIR 2006 Workshop on XML Element Retrieval Methodology.2006.

[9] Dumais, S., Banko, M., Brill, E., Lin, J., and Ng, A.” Web Question Answering: Is More Always Better?” Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland.2002

[10] Fuhr. N., Lalmas. M., Malik. S. and Szlavik. Z. editors. INEX 2004 Proceedings. Springer, 2005.

[11] Fuhr. N., Lalmas. M., Malik. S. and Gabriella Kazai, editors. INEX 2005 Proceedings. Springer, 2006.

[12] Kobayashi M. and Takeda K. “Information retrieval on the web”, IBM Research Report, RT0347, April 2000.

[13] Lin. J., Quan. D., Sinha. V., Bakshi. K., Huynh. D., Katz. B. and Karger. D. R. “The role of context in question answering systems”, In CHI ’03: CHI ’03 Extended Abstracts on Human Factors in Computing Systems, pages 1006–1007, New York, NY, USA, 2003a. ACM Press. 2003.

[14] Maybury, M. Editor. “New Directions in Question Answering”, AAAI/MIT Press, Cambridge. 2004.

[15] Navarro-Prieto. R., Scaife. M. and Rogers. Y. “Cognitive strategies in web searching”. In Proceedings of the 5th Conference on Human Factors & the Web, 1999.

[16] Over. P.”The TREC interactive track: an annotated bibliography”, Information Processing and Management, 37(3):369–381, 2001.

[17] Ogilvie, P. “Retrieval Using Structure for Question Answering.” In the Proceedings of the First Twente Data Management Workshop (TDM`04).2004.

[18] Radev, D., Fan, W., Qi, H., Wu, H., Grewal, A. “Probabilistic Question Answering on the Web. “ Proceedings of the 11th World Wide Web Conference, Honolulu, HI. 2002

[19] Raggett. D., Hors. A. L. and Jacobs. I. editors. “HTML 4.01 Specification”. W3C, 1999. http://www.w3.org/TR/html4/.

[20] Roussinov, D., and Robles-Flores, J. “Web Question Answering: Technology and Business Applications”, In Proceedings of 2004 Americas Conference on Information Systems. August 6 – 8, New York, NY, 2004.

[21] Sigurbjornsson. B. “Focused Information Access using XML Element Retrieval.” University of Amsterdam. SIKS Dissertation Series (nr. 2006-28). 2006.

[22] Sigurbjornsson. B., Kamps. J. and Rijke. M “Focused Access to Wikipedia.” In 6th Dutch-Belgian Information Retrieval Workshop (DIR 2006). Pages: 73-80. 2006.

[23] Slone. D. J.”Internet search approaches: The influence of age, search goals, and experience”, Library & Information Science Research, 25(4):403–418, 2003.

[24] Toms. E., Freund. L., Kopak. R. and Bartlett. J. “The effect of task domain on search”, In Proceedings of CASCON 2003, pages 1–9, 2003.

[25] Tague-Sutcliffe. J., “The pragmatics of information retrieval experimentation, revisited.”, Information Processing and Management, 8(4):467–490, 1992.

[26] Trotman. A. “Wanted: Element retrieval users.” In Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, 2005.