Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Enhanced Architecture for Document Specific Information Retrieval in Dynamic Environment

Sharvan Kumar Garg[*]
Research Scholar, Magadh University
India

R.P.Yadav
Retired Professor & Head, Magadh University
India

Abstract: The foremost problem with document specific web search elucidations is that they are very costly to construct, and necessitate a high level of expertise mingled with a considerable development endeavor. This paper concentrated on approach - specialization by content type. The primary idea of this paper is to construct a framework and tools with which an advanced web developer or consultant can construct a search engine specialization by content type within a small number of man-weeks of endeavor.

Keywords— Metasearch, document-specific, searchengine, recurring, refining.

I. Introduction

Recurring is the action of iterating a process with the objective of approaching a desired outcome. Each recurrence of the process is also called a repetition, and the outcomes of one repetition are used as the initial point for the subsequent repetition. Discriminating presentation or considered manipulation of information to make it more adequate or promising to its beneficiary is referred as refining. It works by employing keywords, and it refines out websites defined by those keywords. All search engines has refine alternatives built in.

There are two most important strategies for constructing document specific information retrieval architectures. This section depicts both in additional detail and addresses their restrictions:

1. Metasearch engine characteristically refers to applications whose objective is to propose wide-ranging searching services along the lines of conventional search engines similar to Bing [1] and Ask [2]. Metasearch engines are among the foremost applications constructed on top of web information retrieval architectures. They dispense their queries to probably several authentic search engines and mingle the outcome to the client. SavvySearch [12] and MetaCrawler [17] were both instigated in mid-1995 and come into view to have been invented autonomously.

2. Document specific search engines with corpus These are autonomous search engines with their own catalog that attain specialization all the way through a cautiously-created corpus (usually by employing *focused crawling* and information mining tools).

Focused crawling is a notion that appears to predate web metasearch. A widespread approach for focused crawling is to utilize a best-first strategy that ranks links according to the match among their anchor text and a model of the search target.

Largely speaking, search can be focused in a number of magnitudes. We are concentrating in this paper on two magnitudes:

- *Topic-specific* concentration.
- *Document-specific* concentration.

Occasionally, information retrieval architecture is focused in both magnitudes, as in the case of the *ResearchIndex* [14], which specializes in scientific research papers (doc-specific). At the same time as certain document types have been planned for a while, it should be distinguished that the more universal facet of document-specific categorization [15] itself has detained additional research interest in recent times, together with work linked to *genre search* [8]. The distinction between *document specific* and *genre* is that document specific refers to the intent of a document, while genre concentrates more on its manner. In the wording both terms are frequently used synonymously. In our work we focus mainly on document specific, not genre.

II. METASEARCH CATEGORIZATION

Based on their intent, we can categorize metasearch engines into

- *All-purpose metasearch engines*
- *Document specific metasearch engines*

The advantage of *all-purpose metasearch engines* (for example, Mamma, http://www.mamma.com) is assumed to be enhanced recall (because the coverage of the metasearch engine is the amalgamation of that of the underlying engines). Though the advantages of metasearch are not restrained to enhanced recall, but also comprise enhanced precision. This is elucidated by the fact that different search engines utilize diverse ranking approaches. For example, if a result constituent has been ranked extremely in several different result lists, it may be certainly a high-quality match.

On the other hand, if an outcome constituent emerges highly ranked in only one or a small number of search engines, probability are that it does not deserve an overall enhanced ranking position. Rank aggregation methods [7] focus on this concern, and as a result construct a fairer and overall additional vigorous ranking with enhanced precision.

The association between a metasearch engine and other search engines may possibly be either one-to-one or one-to-many. We consequently have the following cases:

1. All-purpose metasearch - lying on one search engine
2. All-purpose metasearch - lying on several search engines
3. Document specific metasearch - lying on one search engine
4. Document specific metasearch - lying on several search engines

For all-purpose metasearch we characteristically observe case number two applied (we state case number one in the concern of being complete). In our work we exclusively focus on case number three where we have document specific metasearch lying on one search engine. We plan to widen our work to case number four in upcoming work.

III. OBJECTIVES

An objective for metasearch engines – in addition to higher recall and enhanced precision in the all-purpose case – is to offer elevated precision in the perspective of a specific information necessitate. Moreover, the thought of reusing the large catalog of all-purpose search engines symbolizes a striking strategy for constructing document specific search elucidations. We call these *document specific metasearch engines*. There is an emergent body of work applying document specific metasearch engines to homepage finding [6]. This example of document specific metasearch assists the user exclusively locate personal homepages. Supplementary examples of document specific metasearch engines comprise one that concentrates in the topic of nanotechnology [5], one that discovers news articles associated to closed-caption content [11].

While from an economical viewpoint it appears to be undeniable to use metasearch to construct document specific search elucidations (for example, there is no catalog to construct and to preserve), document specific metasearch engines remain complicated to build up. The major cause for this is because every search engine utilizes its own query language and sentence structure (i.e. syntax), and returns outcomes in some HTML layout (parsing the outcome layout is extremely error-prone). The information requirements of metasearch applications have been so profoundly considered that the STARTS protocol [10] was projected in an endeavor to standardize the interfaces employed by metasearch engines to contact underlying search engines. Such standardization efforts and up-and-coming web service APIs, for example, Bing API [3] will assist to tackle the necessity of standard interfaces to computerize processing. Today's accessible web search engine APIs are targeted further towards human utilization of outcomes, making them not a favored alternative for computerized processing by document specific metasearch applications.

IV. CHALLENGES AND RESTRICTIONS

An additional most important challenge to constructing document specific metasearch engines is how to state and put into practice refining of search outcomes to realize the desired specialization. Nowadays this needs a considerable level of proficiency and manual modification of classifiers, as can seem with examples similar to the homepage finder [13] and work associated to document specific classification [15].

Correspondingly, query formulation that directed to various high excellence outcomes according to a user's document specific information requirement, remains complicated, and there is no readymade elucidation available. Glover et al. [9] are suggesting learning query alterations rooted on support vector machines. The setback with this approach is that it necessitates a considerable training stage and proficiency, but it unquestionably represents a feasible pace towards computerizing query formulation for document specific information requirements. Oyama et al. [16] present a dissimilar approach to query formulation by using decision trees to create keyword spices. This representation does not refine documents, as an alternative, the initiative is to broaden the user's input query with document specific Boolean lexis (keyword spices) to achieve enhanced precision.

Conversely, there are a number of restrictions with this approach. First, even though precision will be elevated with cautiously built queries that contain document specific extensions; our experiments signify that a refining step still appears to be advantageous to weed out unnecessary outcomes. Second, since not every search engines support complete Boolean exploration operators, the outcomes acquired when issuing Boolean queries to these are not precisely defined. Additionally, Oyama et al. [16] do not depicts how the projected techniques can be extended to entirely computerize the building of document-specific information retrieval architecture for a variety of documents.

V. RELATED WORK

One instance of an accepted document specific search engine we stated before is ResearchIndex [14], which is a self-sufficient reference indexing structure for research papers. It employs focused crawling as its key basis to ascertain and produce new information. ResearchIndex [14] also employs web search engines and heuristics to trace research papers. This is done by means of queries for documents that contain definite words (for example, publications, and papers). The acquired search outcomes are then used as a seed list for their crawler. ResearchIndex [14] downloads the research papers themselves, mine meta-data , and do its own cataloging. Presently there appears to more literature on document specific search via focused crawling [4] than on document specific metasearch, although, we consider that the extensive scope of contemporary web search engines will turn this flow. The increased recall attained by focused crawling will be insignificant, while the reduced cost of metasearch will be unquestionable.

Miles Efron [18] proposes generative model-supported metasearch recommends up to date efficiency in data combination, at the same time as also putting the combination operation on firm theoretical balance. Paul Thomas et al. [19] measured a circumstance where one source may perhaps be measured primary, and other secondary, which reveals government metasearch although as well a diversity of enterprise as well as portal settings .He proposes that regardless of the acquaintance of a single catalog, interfaces which support continuing revelation are preferable: especially, it appears that interfaces should maintain users in their selection of source. B.T. Sampath Kumar et al. [20] assess the potential of search engines as well as metasearch engines and demonstrated that no search engines and metasearch engine retrieve extra pertinent information on the World Wide Web .The average accuracy of metasearch engine is far above the ground as contrasted to search engines.

Mohamed Salah Hamdi [21] recommends an information customization structure that mingles metasearch and unmanaged learning. The outcomes retrieved by this system can be extremely pertinent, because it is frequently capturing the foremost items from the relevancy-ranked catalog of hits returned by the individual information retrieval engines. The Kohonen Feature Map is subsequently employed to build a self-managed semantic map such that documents of related contents are positioned near to one another. K.Srinivas et al. [22] proposes a review on different Meta Search Engines and a variety of considerations on which the effectiveness of a Meta Search Engine (MSE) depends. It is implicit that Meta Search Engine demonstrates better presentation than any Search Engine and its performance as well rely on different features similar to recall and accuracy.

Biraj Patel et al. [23] describes metasearch engine a type of means that mingles outcomes of manifold individual search engines by distributing queries to them. In accessible meta search engine, there is no thought of repository of meta search engine. General search engines have their own repository. For rapid communication and effectiveness intention, repository thought should be employed with meta search engine. He confers the representation of repository for meta search engine**.** Biraj Patel et al. [24] confers the requirement of dynamic ranking of information retrieval outcomes in a meta search engine and confers a structure, which permits user to rate information retrieval outcomes. And then subsequent to the meta search engine will revise rank of information retrieval outcomes to create collective outcome according to rating rank. He suggests a novel technique of ranking and information retrieval namely LIKE information retrieval, which permits user to have a selection of rating information retrieval outcomes through interface. And meta search engine retrieves outcomes on the basis of rank rating.

Preeti Naval et al. [25] say that the user make their probe more personalize for the enhanced outcome. Personalization can advance the effectiveness of search engines; augment the amount of outcome as indicated by the user necessity, Offer associate search engine choice on the foundation of user understanding. All these approaches advance information retrieving.

VI. **RRM OUTLINE**

As the preceding sections shows, the metasearch concept of constructing document specific information architecture on top of large, all-purpose web search engines is a widespread one. In this section we introduce a precise approach for document specific metasearch which we refer to as *recurring, refining metasearch (RRM)*. Again, it is our certainty that rising compilation dimensions and web services APIs will direct to amplified concern in this approach.

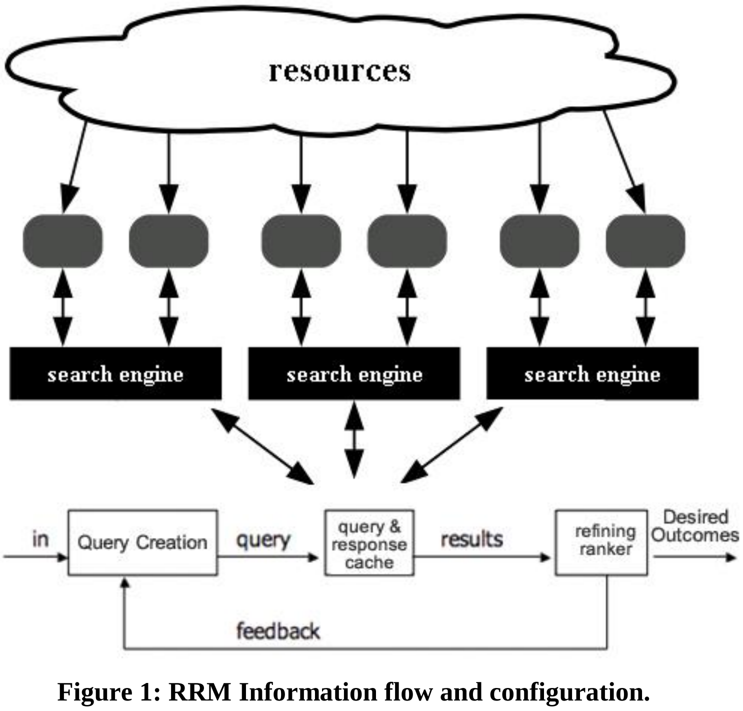

Figure 1: RRM Information flow and configuration.

In the RRM situation several structure of input enters from which outcome is to be produced. Gravano et al. [11], for instance, proposes that input is in the structure of processed closed-description content. He proposes that input is straightforward expressions designating merchandise groups. A catalog of URIs, and potentially several explanatory meta-data, is created as result. Meanwhile this input and outcome, a query-creation algorithm creates queries to be supplied to the information retrieval engine. The information retrieval engine processes the queries and returns outcomes to be supplied into the RRM application's refining and ranking engine (Refiner/Ranker). The refining and ranking engine subsequently chooses and perhaps re-ranks outcomes from the search engine. It also can constructs feedback to the query creation phase, which has an impact on prospective queries supplied to the search engine.

As revealed previously this structure of metasearch is a simplification of conventional metasearch, but the novel components are considerable. In conventional metasearch, the user's query is either passed straightforwardly through to the associated search engines, or is merely to some extent altered for syntactic grounds. In our supplementary broad case, the query creator is additional dynamic, maybe concluding queries, which were not clearly specified [11], or possibly considerably increasing the user's query.

Furthermore, conventional metasearch engines normally re-rank outcomes from the associated search engine. On the other hand, normally they do *not* refine outcomes. In our own understanding, and according to [11], such refining can considerably augment the accuracy of document specific metasearch engines.

VII. Conclusions

The primary idea of this paper is to construct a framework and tools with which an advanced web developer or consultant can construct a search engine specialization by content type within a small number of man-weeks of endeavor. Therefore in this paper we introduced a specific approach for document specific metasearch which we refer to as *recurring, refining metasearch (RRM).*

In the *RRM* perspective some form of input arrives from which outcome is to be created. In [11], for instance, that input is in the form of processed closed-caption content. A catalog of URIs, and probably some explanatory meta-data, is created as outcome.

Amid this input and outcome, a query-creation algorithm creates queries to be supplied to the explore engine. The explore engine processes the queries and returns outcomes to be supplied into the *RRM* application's refining and ranking engine (Refiner/Ranker). The refining and ranking engine then chooses and perhaps re-ranks outcomes from the search engine. It as well can generate feedback to the query creation stage, which has an impact on upcoming queries offered to the search engine.

.

References

[1] http://www.bing.com

[2] http://www.ask.com

[3] Bing Web APIs. *www.**bing**.com/dev/*

[4] Niran Angkawattanawit and Arnon Rungsawang. Learnable crawling: An efficient approach to topic-specific web resource discovery. In *The 2nd International Symposium on Communications and Information Technology (ISCIT2002)*, 2002.

[5] Michael Chau, Hsinchun Chen, Jailun Qin, Yilu Zhou, Yi Qin, Wai-Ki Sung, and Daniel Mc-Donald. Comparison of two approaches to building a vertical search tool: A case study in the nanotechnology domain. In *Proceedings Joint Conference on Digital Libraries, Portland, OR.*, 2002.

[6] Tina Eliassi-Rad and Jude Shavlik. *Intelligent Web agents that learn to retrieve and extract information*. Physica-Verlag GmbH, 2003.

[7] Ronald Fagin, Ravi Kumar, and D. Sivakumar. Efficient similarity search and classification via rank aggregation. In *Proceedings of the 2003 ACM SIGMOD international conference on on Management of data*, pages 301–312. ACM Press, 2003.

[8] A. Finn, N. Kushmerick, and B. Smyth. Genre classification and domain transfer for information filtering. In *Proc. 24th European Colloquium on Information Retrieval Research, Glasgow*, pages 353–362, 2002.

[9] Eric Glover, Gary Flake, Steve Lawrence, William P. Birmingham, Andries Kruger, C. Lee Giles, and David Pennock. Improving category specific web search by learning query modifications. In *Symposium on Applications and the Internet, SAINT*, pages 23–31, San Diego, CA, January 8–12 2001. IEEE Computer Society, CA.

[10] Luis Gravano, Chen-Chuan K. Chang, Hector Garcia-Molina, and Andreas Paepcke. Starts: Stanford proposal for internet meta-searching. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 207–218. ACM Press, 1997.

[11] Monika Henzinger, Bay-Wei Chang, Brian Milch, and Sergey Brin. Query-free news search. In *Twelfth international World Wide Web Conference (WWW-2003), Budapest, Hungary*, May 20-24 2003.

[12] Adele E. Howe and Daniel Dreilinger. SAVVYSEARCH: A metasearch engine that learns which search engines to uery. *AI Magazine*, 18(2):19–25, 1997.

[13] Oren Etzioni Jonathan Shakes, Marc Langheinrich. Dynamic reference sifting: A case study in the homepage domain. In *Sixth International World Wide Web Conference*, pages 1193–1204, Apr. 1997.

[14] Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999.

[15] Katsushi Matsuda and Toshikazu Fukushima. Task-oriented world wide web retrieval by document type classification. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 109–113. ACM Press, 1999.

[16] Satoshi Oyama, Takashi Kokubo, Teruhiro Yamada, Yasuhiko Kitamura, and Toru Ishida. Keyword spices: A new method for building Domain-Specific web search engines. In Bernhard Nebel, editor, *Proceedings of the seventeenth International Conference on Artificial Intelligence (IJCAI- 01)*, pages 1457–1466, San Francisco, CA, August 4–10 2001. Morgan Kaufmann Publishers, Inc.

[17] Erik Selberg and Oren Etzioni. The MetaCrawler Architecture for Resource Aggregation on the Web. *IEEE Expert*, 12(1):8–14, January 1997.

[18] Miles Efron, Generative Model-Based MetaSearch for Data Fusion in Information Retrieval, Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, pp. 153-162, New York, 2009

[19] Paul Thomas, Katherine Noack, Cecile Paris, Evaluating Interfaces for Government Metasearch, Proceedings of the third symposium on Information interaction in context, pp. 65-74, New York, 2010

[20] B.T. Sampath Kumar, S.M. Pavithra, Evaluating the searching capabilities of search engines and metaserach engines:a comparative study, Annals of Library and Information Studies (ALIS), pp. 87-97, Vol.57(2) [June 2010]

[21] Mohamed Salah Hamdi, SOMSE: A semantic map based meta-search engine for the purpose of web information customization, Applied Soft Computing, Volume 11, Issue 1, pp. 1310–1321, 2011

[22] K.Srinivas, P.V.S.Srinivas and A.Govardhan, A Survey on the Performance Evaluation of Various Meta Search Engines, IJCSI International Journal of Computer Science Issues, pp. 359-364, Vol. 8, Issue 3, No. 2, May 2011

[23] Patel, Birajkumar; Shah, Dipti, Incorporation of Databases for Faster Meta Search Engine, International Journal of Advanced Research in Computer Science . Vol. 3 Issue 6, pp. 209-210, Nov/Dec2012

[24] Biraj Patel, Dipti Shah, LIKE Search on Meta Search Engine, International Journal of Advanced Research in Computer Science and Software Engineering, pp. 259-362, Volume 3, Issue 6, June 2013

[25] Preeti Naval, Priyanka Singh, a Survey on Personalized Meta Search Engine, International Journal of Advanced Research in Computer Science and Software Engineering, pp. 322-325, Volume 2, Issue 3, March 2012