IJCSC 0973-7391 Volume 7 • Number 1 Sept 2015 - March 2016 pp.100-105 available online at www.csjournalss.com

Aggregated Search-A rising need for Information Retrieval

Aashka Kotecha[1], Shailee Patel[2] Shivani Desai[3]
[1]Student (13MCEN06), Dept. of Computer Science and Engg.,Nirma University, Ahmedabad, India
[2] Student (13MCEN20), Dept. of Computer Science and Engg.,Nirma University, Ahmedabad, India
[3]Assistant Professor, Dept. of Computer Science and Engg.,Nirma University, Ahmedabad, India
13mcen06@nirmauni.ac.in, 13mcen20@nirmauni.ac.in, shivani.desai@nirmauni.ac.in

Abstract: A regular search results into ranked pages based upon various algorithms and their previous hits, this may not be advantageous all the time. Thus, information required by user will be scattered in various documents and user will require manually combining them. Aggregated search does the work of combining the information from different sources. Aggregated search will result with images as well as multimedia for a query if related information is found. Aggregated search comes in advance search engine mechanism with various facilities which we tried to explore in this paper. This paper also proposes a viable solution to check the credibility and trustworthiness of the data source used in data aggregation.

Keywords: Aggregated search; Information retrieval; Solution to Aggregated Search; Relational Search; Natural Language Processing; Federated Search; Credibility; Trustworthiness.

I. INTRODUCTION

For a simple query there may be thousands of relevant results over web, but to specifically find some information over multimedia we need to specify parameters of search. Also it is not possible for even the perfect search engine to retrieve precise information if it is not clubbed in database. Thus at the end we need to combine results for gathering the perfect information for user. Aggregated search does this job of combining information.

A. Aggregated Search Techniques[1]

Aggregated search consist of basically two searching techniques *Cross vertical Aggregated search (cvAS)* and *Relational Aggregated search (RAS).* These two techniques can be understood as:

1) *Cross-vertical Aggregated Search*: This technique involves the multimedia information retrieval for a query if the content is matched for videos, images, etc. It is child technique of *Federated Search* which allows simultaneous searching for content by distributing query over search engines and then will aggregate the results by matching their common themes. It is most common for all search engines to have this technology as there subpart in their search engine. Vertical searching for information is quite basic requirement for every search engine.
2) *Relational Aggregated Search:* Relational aggregated search helps to find dependences among various keywords of query. For example Delhi as capital of India whilequery term may only be Delhi, while searching for ChetanBhagat it retrieves few famous books written by him, and similar relations can be searched or retrieved in this search. Basic search will contain information from WIKIPEDIA and key words related to paring of terms previously searched for this keyword.

   Figures (Fig 1, 2 and 3) shows various results for single query of Manmohan Singh , retrieved information will vary as per search engine, but all the search engines retrieves web links as well as Images for Manmohan Singh and videos related to him. Wikipedia information is also shown side by just to have quick glance without opening the site we can have overview of the results so that we don't need to open the whole web link and required information can obtained.

Aggregated search targets information parts related to each other whether in form of images, videos or web links. Aggregated search will combine the results by crawling through web pages through keywords. There are two files for any search link as humans.txt and robots.txt, which will help keywords to be retrieved, searched faster than regular search. [1]

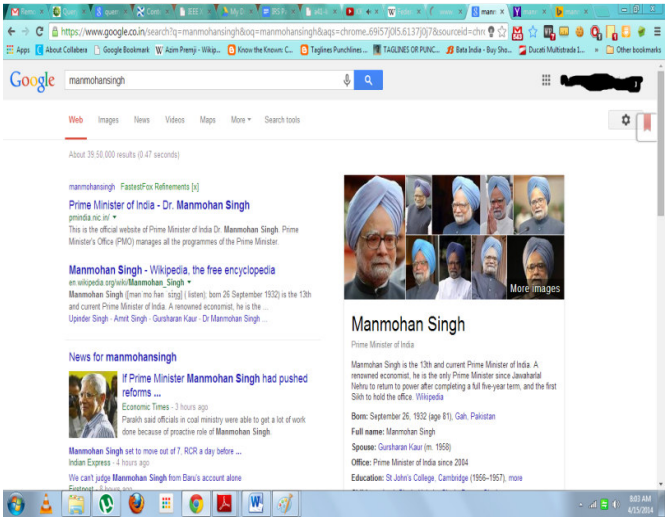

Fig.1 Google search for our Prime Minister Manmohan Singh

Fig 2. Yahoo Search for Manmohan Singh

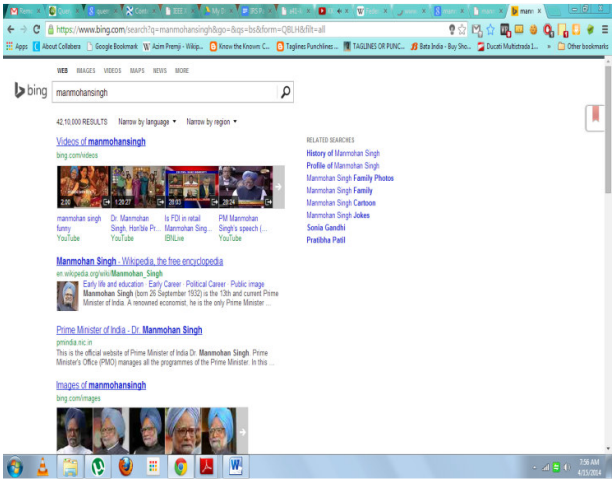

Fig 3. Bing results for Manmohan Singh

B. Problems

The major problem with aggregated search is the credibility of the source that is presented to the user. In more simplified manner we can say the source from where the data gets aggregated needs to be trustworthy and verified.

The basic problem with aggregated search will be that all application will not be applicable to map with data over web, as we cannot predict the relation and no specific rule for having relations between information about search query.

Aggregated search cannot be built prior as they cannot be predicted due to large variety of data available over web. Thus we need dynamic approach to resolve dependency for keywords. Runtime our algorithm will decide which information we will be clubbing for a user based upon his/her query.

This paper is divided into 3 sections. First section describes requirement behind implementing aggregated search. The second section discusses general framework for decomposing queries, the third section has Analysis of results based on user feedback and the last section has the solution to the source trustworthiness problem.

II. NEED FOR AGGREGATION

Most of algorithms work on ranking system whether it may be Boolean, vector or user feedback. It results into ranked list of web-pages related to query. Now in most cases a user may require only glimpse of information which he/she want. Thus as ranking only results web links user will need to provide every time file format to get a particular file. Few of problems with ranking retrieval system are: [1]

A. Data Dispersal: The relevant information is scattered over various documents and user need to go through every document for his query manually. This can be proved time consuming as well as cumbersome to process.

B. Lack Of Focus: The results show directly the part of document which matches the query, while it may not be necessarily relevant to required document. The title, link to documents, the web reference etc. are also important entities thus, we also need to aggregate those results with various other file formats where we link multimedia with its tags and keywords.

C. Ambiguity[5]**:** Many queries may prove to ambiguous as query term will mean different for different synonyms. For example Orange, it can mean color or fruit or mint which can mean flavor, plant or Linux OS. This it is quite necessary to relate the search results across web-pages, ideally search engine should retrieve only one answer for particular query interpretation but aggregation can play important rule if user is totally unaware of contents of query so that he can find results as per his context.

D. Non-Uniformity: Many a times results searched are non-uniform and fetched from non-trust worthy sources, and retrieved only because the query terms are matched thus it may not be advisable to retrieve results based only upon keywords, search engine also need to have cookies storing facility so that previous results and area of interest of particular user can be mapped and ontology map can map the results related to query keywords.

III. GENERAL FRAMEWORK FOR AGGREGATION

In aggregation basically we divide the information into nuggets. Now this nuggets are pointing to various keywords as per their literal meaning and various nuggets will be formed for every different keyword for example nugget for ORANGE color and nugget for ORANGE fruit will be different.

In aggregation basically we divide the information into nuggets. Now this nuggets are pointing to various keywords as per their literal meaning and various nuggets will be formed for every different keyword for example nugget for ORANGE color and nugget for ORANGE fruit will be different.General framework would consist of procedure to accumulate results of query and displaying them over predefined interface in a particular format.

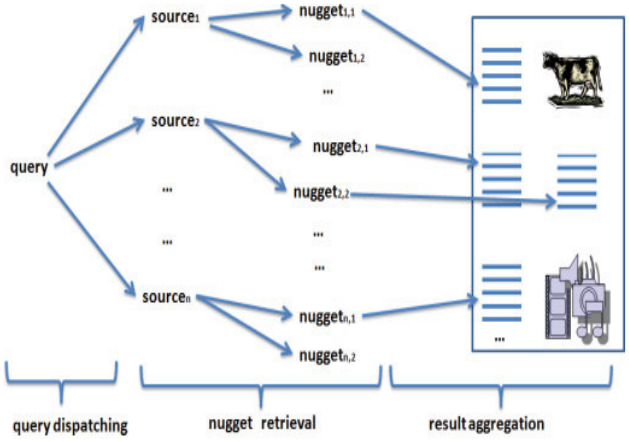

Fig 4. Generalized flow

The figure shows various parts of aggregated search as Query Dispatcher (QD), Nugget Retrieval (NR) and Result Aggregation (RA).

Query Dispatcher: it takes care of how a query should be dispatched over a database by preprocessing the contents of query. It will start relating the terms from the keywords for example if we search any particulars name his/her personal information will start gathering from his/her birth date, place to current working positions with all biography information placed over whole web servers. Query reformulation will take place and various knowledge bases will be checked for related information.

Nugget Retrieval: [11] There are lots of techniques for nuggets retrieval mostly based upon ranked sets of documents. This includes meta-searching [3], federated searching [2] and Mashups with cross-vertical search techniques [4].

Result Aggregation: Result aggregation contains various methods as follows: [1]

1.Sorting - All nuggets will be sorted based on particular parameter like time of upload , size of document , author , location , popularity etc. Although it is different from ranked document retrieval but it incorporates some of parameters as its subpart.

2.Grouping - Grouping will be done of various nuggets based upon particular property similar to ontology so that proper clustering [6] and classification [7] can be done for information retrieval

3.Merging - Merging is also important part as grouping as aggregation and decomposition of data both are of same importance as one can't be done without other as reverse process. As it takes multiple documents and produces a multi document summary.

4.Splitting - It is basically opposite of merging. Decomposition can be proved beneficial as from decomposition we can separate various HTML tags based on their classification and also XML data can be separated for various formats.

5.Extracting - this refers to extracting synonymic information retrieval of keywords so that user can select his context from all the possible outcomes so as to provide versatility and flexibility for user and developer.

IV. ANALYSIS OF THE SEARCH APPROACHES

It incorporates various properties so as to aggregate resulting data for a particular information based on various IR systems like natural language processing, question-answering , federated search , Mashups, cross-vertical aggregated search etc.

A. Natural Language processing [8]: Basic usage of Natural Language processing is to relive user from burden of going through all documents manually. For a web search data source is whole network of servers including all public domains and reports etc. Through natural language generation we can extract type of information we want , for e.g. Delhi as search query may result into map of delhi , path from current location to delhi , history of delhi , current news related to delhi , temperature of Delhi etc. Thus NLG plays important role while processing queries.

B. Question Answering : For a regular search results are kept just as documents containing keywords but for aggregated search we can assemble various documents through particular datasets like popular sites of

Wikipedia , new sites like AajTak etc. for given query. This follows 5Ws pattern where data is extracted based on question like “who”, “what”, ”where”, “when” and “why” this answers are searched and data is retrieved. This task is usually not easy as we have to implement machine learning and neural networks for retrieving related data and assemble them at single interface designed for data retrieval.

C. **Federated Search**: [14] It is also known as distributed information retrieval from various distributed sources. In federated search, our search engine parallel fires query on various sources and on various data types so that heterogeneous set of results are obtained. At the end it assembles the result over interface so that user get ease of finding related data according to his needs.

D. **Mashups**: Mashups are interesting tool where search engine assembles data as well as services of various kinds so that a new service can be provided to user. Yahoo Pipes is one of most popular mashup available.

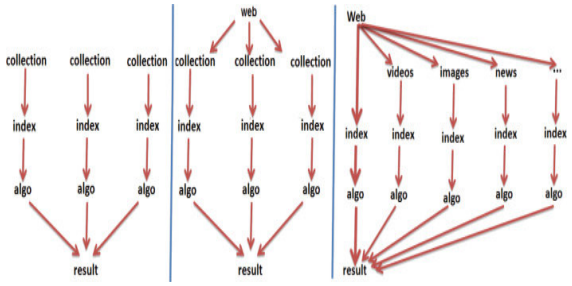

Fig 5. schemas for federated search (left), meta-search (center), and cross-vertical aggregated search

E. **Cross-Vertical Aggregated Search**: It’s the task of scheduling and assembling information from vertical search and web search, usually done in Web search content.

V. PROPOSED SOLUTION

To establish a level of credibility of the information in query online users incline more towards both trustworthiness and expertise cues.[9] In this context, user would verify as to who wrote some information to assess and whether the author as well as the authority are trustworthy or not. If this is the case then strong credentials and objectivity would suffice the need. But, the major problem with this is that a great deal of online information is detached from this credential and authority cues. In particular user created content platforms like Wiki, Review and rating websites; Blogs, twitter etc. provide very minimal direct cues of expertise.

Thus, the social element attached to the information is a key element when users evaluate its credibility. [10]

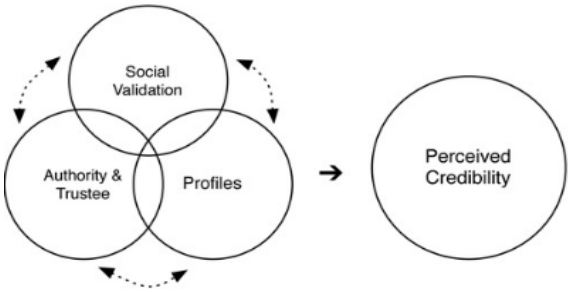

Fig. 6 Example of Aggregated trust worthiness

The above figure shows an approach to attain credibility and trustworthiness of an information source via social element.[13]

A. **Perceived credibility:** It represents the degree to which a user believes the information presented to him/her.

B. **Social Validation:** It includes large scale verifications made by others for e.g. comments, Facebook Likes, Shares, Social bookmarks, ratings etc. It may include profiles but not constrained to them. In a nutshell social validation simply means more the people acknowledge a certain piece of information the more trustworthy it is.

C. Profiles: It provides the baseline for online identity as well as base for evaluation. For e.g. LinkedIn profile, Twitter stream, personal website or blog. While we are accessing some vital information it is very critical to have a known identity and profile.

D. Authority and Trustee: It includes known brands and authority on the matter. For e.g. New York Times, Stanford University etc. but also Trustees verifying lesser known sources e.g. Social Network Friends, Wikipedia references, Twitter Personas etc.

The above factors are interrelated and interdependent on each other. This model demonstrates how social validation provides verification of an authority which in turn may provide verification of a specific profile focusing our evaluation process and establishing perceived level of credibility of information. [12]

VI. CONCLUSION

This paper suggests approaches for query matching and result aggregation providing the user with an aggregated search. This approach makes use of relations to assemble and retrieve search results, also enabling retrieving at finer granularity and composing new objects of related content. The papers also provide solution to a major problem prevailing with aggregated search namely authorization and verification of source of contents used in aggregation.

VII. REFERENCES

[1] *Aggregated Search: A New Information Retrieval Paradigm.* ARLIND, KOPLIKU, KAREN, PINEL-SAUVAGNAT and MOHAND, BOUGHANEM**.** 3, article 41, s.l. : ACM Computing Surveys, January 2014, Vol. 46.

[2] Callan, Jamie**.***Distributed information retrieval. In Advances in Information Retrieval.* Dordrecht : Kluwer Academic Publishers, 2000. 235-266K

[3] *Multi-service search and comparison using the MetaCrawler.* Erik, Selberg and Oren, Etzioni. s.l. : In Proc.of the 4th International World Wide Web Conference, 1995. 195-208..

[4] *A methodology for evaluating aggregated search results.* Jaime, Arguello, et al**.** s.l. : In Proc. of ECIR 2011, 2011. 141-152.

[5] Karen, Sarck-Jones, Robertson, Stephen E. and Sanderson, Mark. Ambiguous requests: Implications for retrieval tests, systems and theories. *SIGIR Forum.* 2007, Vols. 41,2, 8-17*A methodology for evaluating aggregated search results.* Jaime, Arguello, et al**.**s.l. : In Proc. of ECIR 2011, 2011. 141-152.

[6] *A methodology for evaluating aggregated search results.* Jaime, Arguello, et al. s.l. : In Proc. of ECIR 2011, 2011. 141-152.

[7] Christopher, Manning, Prabhakar, Raghavan and Heinrich, Schutze**.***Introduction to Information Retrieval.* s.l. : Cambridge University Press, 2008.

[8] *Focused and aggregated search: A perspective from natural language generation.* Cecile, Paris, Stephen, Wan and Paul, Thomas**.** 3, 2010 : Information Retrieval Journal, Vol. 44.

[9] *The elements of computer credibility.* Fogg and Tseng, Hsiang**.** s.l. : CHI '99: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1999. 80–87.

[10] *Persuasive technology: Using computers to change what we think and do.* Fogg**.** Boston : Morgan Kaufmann, 2003.

[11] NuggetMine: Intelligent groupware for opportunistically sharing information nuggets. Jeremy, Goecks**.**s.l. : In Proc. of IUI 2002, 2002. 87–94

[12] Lankes, R. David**.** Trusting the Internet: New approaches to credibility tools. *Digital media, youth, and credibility.* 2008, 101–122.

[13] *Aggregated Trustworthiness: Redefining online credibility through social validation.* Johan, Jessen and Anker, Helms,Jorgensen**.** 1-2, s.l. : first monday, January,2012, Vol. 17.

[14] *Leveraging query association in federated search.* Pal, Aditya and Kawale, Jaya**.** s.l. : In Proc. of SIGIR 2008 Workshop on Aggregated Search., 2008