ADVANCE INFORMATION RETRIEVAL STANDARD: A SURVEY

Dhara Bakrola[1], Prof. Snehal Gandhi[2]
[1,2]Sarvajanik College of Engineering and Technology, Surat, India
Email:[1]bakrola.dhara@gmail.com, [2]snehal.gandhi@scet.ac.in

Abstract— The World Wide Web (WWW) is a substantial source of information which is growing rapidly. Many organization and companies are emerging application on web environment. Search engine is information retrieval system which is design to rummage information over the WWW. Commercial search service provider such as Google, Bing, and Yahoo provide access to a wide range of specialized services that are called verticals. When information is not enough, relevant information is scattered in different types of documents like image, video, news, etc. Therefore, aggregated search is introduced that is an effective approach for blending vertical in single result page. The main goal of aggregated search is to satisfy user's need by searching and assembling information from variety of verticals and placing them into a single result page. This paper contain exhaustive survey of the current evaluation on advance information retrieval, Pre-retrieval and Post-retrieval features of aggregated search. We have come across some limitations such as: vertical selection, vertical presentation, and result representation.

Index Terms— Aggregated Search, Information Retrieval, vertical search.

I. INTRODUCTION

The World Wide Web (WWW) is an incredible source of information, it can be considered as substantial distributed information system which provides access to shared data items through Information Retrieval system [1]. There are thousands of different search engines available with their own abilities and features. It contain various information related to education, news, scientific research, sports, stock exchange, entertainment, map, weather, shopping etc.

The main goal of a search engine is to display available links which are relevant to the query fired by the user. In recent development, search engines have elongated their services to include search (called as vertical search), on specialized collections of documents that are called as verticals, which focuses on specific domains (e.g., travel, news, shopping) or media types (e.g., blog, image, video) [2], [3]. Users believe that relevant information exists in a vertical and may submit their queries directly to a vertical search engine. Users are unaware or not willing to use a suitable vertical and therefore they submit their queries directly to the "*general*" web search. Users may unaware of a suitable vertical, or simply not willing to search a specific vertical, user would submit their queries directly to the "general" web search engine. However, users who type certain queries, for instance, "sports bikes", may actually be interested in seeing images of sports bikes even if they did not submit this query to an image vertical search. Search engines include suitable results from relevant vertical within the "standard" web results. This referred to as aggregated search and has now been employed by commercial search engines such as Google, Bing, and Yahoo.

Aggregated search endeavors to achieve diversity by presenting search result from

different source of information called as verticals (e.g., image, video, news, blog, etc.) and present them with standard web result on single result page [4]. This comes in distinction with the common search standard, where users were made available with a list of information sources, and they have to scrutinize on term, to find relevant content.

This paper is further structured as follows: Section II represents the motivation towards aggregated search. Subsequent section III describes theoretical background and survey and section IV illustrates various issues of aggregated search. We have put forward section V discusses the conclusion of the study.

II. MOTIVATION

If Most of the information retrieval systems providing information according to the rank retrieval (rank list) model that means, documents that matches to the requested query are returned as a ranked documents to the user. Generally, query response are arranged in a ranked on the basis of some scoring function. Scoring function combines diverse characteristics produced by the documents and query. But there are some constraints [1], [2], [3] and [5] of conventional information retrieval system which are as follows:

- Rank retrieval [1]: Results are represented according to their rank.
- Scattered data [2]: For particular query single document is not enough but related results are available on different document type (relevant documents are scattered). In such cases user has to fulfill their need by finding information in different documents.
- Short of focus [3]: For each user issued query, information retrieval system has to return related documents(according to the ranked list).For web search results instead of serving whole document as response of query just provide a part of document.
- Vague query [1], [2] and [5]: Queries which have more than one meaning called as ambiguous queries. The reference example is Saturn which can be referred as it is a sixth planet from the sun, it is an Operating System, and also a car company. Ultimately, it should return single answer for each query interpretation and it can be several ranked lists or related sets of results.

These limitations are overcome by a novel paradigm called as aggregated search. In aggregated retrieval [5] there is no compulsion that final result should be the rank result, it can be any arrangement of useful content (or document) which is fruitful to find necessary information for the user requested query.

III. THEORETICAL BACKGROUND AND SURVEY

The goal of aggregated search is to provide integrated access to all these different sources from a single page. From a system perspective, this task is divided in to two part: First, Predicting verticals to present results (called as vertical selection) subsequent, predicting where to present them with web results (called as vertical presentation) [4]. Generally, selected vertical is blended with few of its top results and presented somewhere above, below, within the first page of the web results.

A. Archetypal Framework of Aggregated Search

The clear idea of aggregated search was initiated as a universal search by Google in 2007. A. Kopliku et al. [1] have proposed a framework for aggregated search which provide the study of various approaches which are related to aggregated search. This framework contain three main components and they are: Query Dispatching (QD), Nugget Retrieval (NR), and Result Aggregation (RA) (see Fig. 1). It means that each requested query processed with all the term, Query will be dispatched in several sources in order to find relevant information for each source and return such information. All results are assembled in final result and answers can be organized irrespective of query sense. This makes a clear difference of result aggregation process from existing approaches.

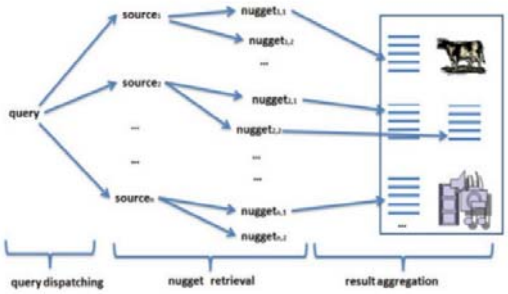

Fig. 1. Archetypal Framework for Aggregated Search [1]

- *Query Dispatching*: It is an initial steps that proceed towards query matching and that will decide the solution for a given query.
- *Nugget Retrieval*: It is an intermediate step which correspond with the meaning of a source which takes input as a query and matches with the relevant information nuggets (nuggets is a piece of valuable information).
- *Result Aggregation*: It starts when it will have possible relevant set of information nuggets. It involve different way of gathering content by applying some actions (aggregation actions) on search result before presenting to the user and these actions are shorting ,grouping, merging, splitting, and extracting. These actions can be performed alone or in any combination.

The different search engine provides different specialized services called as verticals and its brief description is shown in Table 1.

B. Exploration of Aggregated search Approach

This section describes various types of aggregated search approaches. First of all, there are some approaches in natural language generation that focusses on result aggregation. Then we have question answering approach in which there are inspiring case studies with respect to query interpretation and result aggregation. On the other hand there two prominent approaches such as Cross Vertical Aggregated search and relational aggregated search which are described as follows:

- *Cross Vertical Aggregated Search*: Cross-vertical aggregated search [6], [7] and [8] is a task of diversifying search results with different vertical standards on a single result page. The advantage of cross-vertical aggregated search is that the relevant verticals are represented in blended manner within a single result page. In cross-vertical aggregated search, it will search relevant result in each vertical for each user issued query. If relevant information found in specific vertical than result are gathered and represented in search engine result page. This approach can be considered as "divide and conquer" approach.
- *Relational Aggregated Search*: A. Kopliku et al. [9], [10] have developed a fretwork which consider relations between results called as relational aggregated search (see Fig. 3). Relational aggregated search focuses on relation for example, consider a query "Us President" than relational aggregated search provide relational result (date of birth, date of death, achievements, etc.) related to that query (See Fig. 3). Examples of relational aggregated search are Google Squared, Wolfram Alpha. It requires more focus on precision and recall and these can be a future direction for researcher.

Fig. 2. Example of Cross-Vertical Aggregated search

Table 1. Vertical Description

Vertical	Description
Images	Retrieves online images.
Videos	Retrieves online videos.
Finance	Retrieves information related to economics data and business.
Movies	Retrieves movie shows times.
Music	Retrieves musician profiles.
News	Retrieves News articles.
Health	Retrieves information related to the health (articles)
References	Retrieves entries of Encyclopedia.
Travel	Retrieves reviews of travel and accommodation.
Job	Listing of job.
Games	Retrieves online games.
Local	Business listing.
Map	Retrieves maps and direction.
Shopping	Reviews and product listing.
Sports	Statistics, articles and scores of sports.
TV	Listing of television

Fig. 3. Example of relational aggregated search

C. Literature Survey

In this exhaustive literature survey on aggregated search for different verticals, we have found that aggregated search is evaluated on the basis of vertical position [2], understanding user's sequence of action and simple action [3], predicting user's preferences on web vertical [7], effect of thumbnail and spillover effect [9], vertical position on different page [10], position of result [8] such as image, video, news, vertical presentation[11] of image, video, news, blog etc. user's decision process on web [12], for improving the ranking results of web search [6] by considering dwell time and clicks. We found most promising and novel approach in literature is cross-vertical aggregated search and relational aggregate search described in section III.

Classification and ranking algorithm can use the features that can be turned from evidence and these features are divide into pre-retrieval and post retrieval features. Pre-retrieval features are generated before issuing the query and post-retrieval features are generated after issuing the query [13]. Various pre-retrieval and post-retrieval features are- shown in Table 2.

Table 2. Classification of Features on the basis of their Retrieval

Pre-retrieval Feature [1][8][15]	
Features	**Description**
Named-entity	Named entity feature shows the existence of named entities of several type in the query.
Click through investigation	These features are created from the documents that user have been clicked for query and click can be considered as implicit feedback.
Vertical intent	Various phrase point toward a query intent such as Video, Image, etc. Most of the time they provide query intent as well as source intent.
Category Representation	With the help of classification of query into pre-defined classes such as technology , music, etc. these feature can be generated
Post-retrieval Feature [8][15]	
Features	**Description**
Identical Match Score	This feature Corresponds to equivalent score on the results calculated

Pre-retrieval Feature [1][8][15]	
Features	**Description**
	consistently across various sources.
Source relevance score	This feature is measure of relevance of the different sources for user issued query.
Quantity of result	This feature is a count of result retrieved from different sources.

IV. Issues of Aggregated Search

Aggregated search is an effective approach of presenting the search results from diverse sources in to a single result page. But there are some critical issues associated with it which are described as follows:

A. Vertical Selection and Vertical Presentation

Task of selecting the relevant vertical for each user issued query is called as vertical selection or source selection. Majority of the search engines are source selective in order to avoid a long query response time. Vertical selection is one of the familiar problems in aggregated search with multiple verticals. Its goal is to select the vertical that are likely to answer the issued query. While selecting a vertical: what are the Key term required to identify query intent? , how many features need to be consider and how to represent them in terms of feature? , which vertical to be used? are the unresolved questions. Aggregated search deals with extremely heterogeneous it requires large access time than conventional system. Another side in vertical representation sources have internal representation in cross-vertical aggregated search. Representation of vertical should be in text description or in terms of some feature? How to assemble result that are coming from different verticals and how to represent them?

B. Result Representation

Result aggregation concern with the ranking of results whereas, result presentation deals with presentation interface. One way to present vertical result is contents of the same are placed in a single predefined panel is called unblended interface [4]. Another way of presentation is combine search result from different verticals with one another is called blended interface.

J. Arguello et al. [2] and M. Lalmas et al. [4] have described two type of result page design in aggregated search. One, in which results from different sources are blended onto a single interface (called as blended design) and another, in which result from the different source are presented separately in panel (called as non-blended design).Google as a universal search has applied blended design and nowadays, many other search engines are serving blended design to the user. In aggregated search blended result of a same vertical are presented in a slot. In blended design, the main ranking criteria within verticals and athwart verticals are considered. Results from the same vertical are slotted together and each result raked with respect to their expected relevance to that query nevertheless, whole slot is ranked with respect to one another.

M.Lalmas et al. [4] defined non-blended design, and results from each vertical represented separately in panel. Panel is also referred as tile in search engine terminology. Examples of non-blended design are alpha yahoo, Kosmix and Naver. Usually, web search results are displayed in a left side and different panel has no relationship with one another moreover placement of different panels are predefined.

Fig. 4. Example of blended Design

Fig. 5. Example of Non-Blended Design [15]

In aggregated search each source will perform nugget retrieval process and retrieved nuggets will have to be assembled in one result page.

Though the task are distributed, the problem is far away from being solved. It is not easy to decide which source would be used and how the retrieved information (result) should be collected and presented. We enlist some of the major issues which have the consideration of current research.

- Vertical selection and representation: which vertical should be used? How should they be represented in terms of features?
- Aggregation of result: How to assembled search results from different vertical?
- Result presentation: which are the acceptable interface for Cross-Vertical Aggregated Search? And how to arrange vertical in interface? A main challenge is to identify the best position to place item retrieved from relevant verticals on the result page to maximizing click through rate (e.g. blended design or non-blended design)

C. Evaluation of user's judgment

Information retrieval system is designed to find information over the WWW. Information which is available on World Wide Web is only for users. Information retrieval system help user to find requested information over the WWW by providing relevant, essential, structural information within fraction of time. User's relevance measure is essential and it will assist to improve search engine performance. Search relevance can be measured by human evaluation and judgments [15]. Human judgment or feedback (or ratings) can be obtained in two way, one is explicit and another is implicit feedback they are discussed below.

- Explicit ratings of humans are expensive to obtain since, millions of user interacts with Web and to get explicit feedback from each it is quite difficult task.
- In this method system keep asking user to rate the document they have visited. User may stop providing explicit feedback, when he will not found beneficial [15]. In explicit feedback user inform directly to the system what they feel about a part of information which affect normal prototype of browsing.
- Implicit ratings of a human can be used to obtain a large quantity of information for maintain, evaluating, and improving information retrieval (IR) system. Implicit feedback reduces the cost of user inspection [16]. This method can take contribution of each user and can be used in large scale operational environment. Implicit feedback is the normal activities which are generally performed by the user while browsing the Web such as mouse clicks, mouse movements, bookmarking, dwell time, query reformulation, eye-tracking, keyboard activities through which we can identify user interest and its interaction with Web. We will use implicit feedback to identify which vertical is useful for what type of query (for issued query).

In order to improve search engine performance user's feedback is essential. It can be explicit or implicit described above.

V. CONCLUSION

This survey shows that there are numerous approaches available that go beyond query matching with an additional effort on result aggregation. Aggregated search is a novel paradigm of information retrieval which present the search results from diverse sources and present them in a single interface. Cross-vertical aggregated search is a very prevalent type of aggregated search which produces diverse search results by searching and assembling relevant information from variety of sources and presenting them on a single result page. We had presented an exhaustive literature survey highlighting various pre-retrieval and post-retrieval features of cross-vertical aggregated search. We had also accentuate various issues of cross-vertical aggregated search such as vertical selection, vertical representation and result presentation. We believe that this survey in combination with other ongoing research indicate that upcoming IR can assimilate more focus, structure and semantics in search results.

REFERENCES

[1] A. Kopliku, K. Pinel-Sauvagnat, and M. Boughanem, "Aggregated Search: A new Information Retrieval Paradigm," Presented

at ACM Computing Survey, USA, pp. 41-41,2014

[2] J. Arugello, R. Capra, and W. Wu, "Factors Affecting Aggregated Search Coherence and Search Behavior," In Proc. 19th Int. Conf. Information and Knowledge Management, USA, pp. 1989-1998,2013

[3] J. Arguello, R. Capra, "The Effect of Aggregated search Behavior," In Proc. 21th Int. Conf. Information and Knowledge Management, USA., 2012, pp. 1293-1302.

[4] M. Lalmas, "Advance topic in Information Retrieval," Springer Berlin Heidelberg, pp. 109-123,2011

[5] D. Jiang, J. Pei, and H. Li, "Mining Search and Browse Logs for Web Search: A Survey," Presented at ACM Trans. , USA., pp. 57-57, 2013

[6] Rodrygo L. Santos, C. Macdonald, and L. Ounis, "Advance in Information Retrieval theory," Springer Berlin Heidelberg, pp. 250-261, 2011

[7] J. Arguello, F. Diaz, J. Callan, and J. Crespo, " Source of Evidence for Vertical Selection," In Proc.19th Int. Conf. Information Retrieval, USA, pp. 315-322, 2009

[8] F. Diaz, M. Lalmas, and M. Shokouhi, "From Federated to Aggregated Search," In Proc. 33rd Int. ACM SIGIR Conf. Information Retrieval, USA, pp. 141-152,2011

[9] A. Kopliku, K. Pinel-Sauvagant, and M. Boughanem, "String Processing and Information Retrieval," Springer Berlin Heidelberg, pp. 117-128, 2011

[10] A. kopliku, K. Pine-Sauvagnat, and M. Boughanem, "Retrieving attributes using web tables," In Proc. 11th Annu. Conf. Digital libraries, USA, pp. 397-398, 2011

[11] J. Arguello, F. Diaz, J. Callan, and B. Carterette, "A Methodology for Evaluating Aggregated Search Results," In Proc. 33rd Euro. Conf. advance in information retrieval, Heidelberg, pp. 141-152, 2011

[12] J. Arguello, F. Diaz, and J. Paiement, " Vertical Selection in the Presence of Unblended Verticals," In Proc. 33rd Int. ACM SIGIR Conf. Information and Knowledge Management, USA, pp. 691-698,2010

[13] J. Arguello, F. Diaz, and J. Calan, "Learning to Aggregate Vertical Results into Web Search Results," In Proc. 20th Int. Conf. Information and Knowledge Management, USA, pp. 201-210, 2011

[14] M. Claypool, P. Le, M. Wased, and D. Brown, "Implicit Interest Indicators,'' In Proc. 6th Int. Conf. Intelligent user interfaces, USA., 2001, pp. 33-40.

[15] Z. Dou, R. Song, X. Yuan, and J. Wen, "Are Click-through Data Adequate for Learning Web Search Rankings?," In Proc. 17th Int. Conf. Information and Knowledge Management, USA, pp. 73-82,2008

[16] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, " Accurately Interpreting Click-through Data as Implicit Feedback," In Proc. Annu. Int. Conf. Research and Development in Information Retrieval, USA, pp. 154-161, 2005