Aggregated Search

Mounia Lalmas
mounia@acm.org

Abstract. To support broad queries or ambiguous information needs, providing diverse search results to users has become increasingly necessary. Aggregated search attempts to achieve diversity by presenting search results from different information sources, so-called verticals (image, video, blog, news, etc), in addition to the standard web results, on one result page. This comes in contrast with the common search paradigm, where users are provided with a list of information sources, which they have to examine in turn to find relevant content. All major search engines are now performing some levels of aggregated search. This chapter provides an overview of the current developments in aggregated search.

Keywords: vertical, vertical search, vertical selection, vertical representation, result presentation, diversity, universal search

1 Introduction

A main goal of a web search engine is to display links to relevant web pages for each issued user query. In recent years, search engines have extended their services to include search, so-called *vertical search*, on specialised collections of documents, so-called *verticals*, focused on specific domains (e.g., news, travel, shopping) or media/genre types (e.g., image, video, blog). Users believing that relevant content exists in a vertical may submit their queries directly to a vertical search engine. Users unaware of a relevant vertical, or simply not wishing or willing to search a specific vertical, would submit their queries directly to the "general" web search engine. However, even when doing so, users who type certain queries, for instance, "red cars", may actually be interested in seeing images of red cars even if they did not submit this query to an image vertical search. To address this, search engines include, when appropriate, results from relevant vertical within the "standard" web results. This is referred to as *aggregated search* and has now been implemented by major search engines.

Aggregated search addresses the task of searching and assembling information from a variety of information sources on the web (i.e., the verticals) and placing it in a single interface. There are differences between "standard" web search and aggregated search[1]. With the former, documents of the same nature

[1] Although now standard web search is mostly aggregated search. Standard here refers to the pre-aggregated search era.

are compared, e.g. web pages or images, and ranked according to their estimated relevance to the query. With the latter, documents of a different nature are compared, e.g. web pages against images, and their relevance estimated with respect to each other. These heterogeneous information items have different features, and therefore cannot be ranked using the same algorithms. Also, for some queries (e.g. "red car"), it may make more sense to return, in addition to standard web results, documents from one vertical (e.g., image vertical) than from another (e.g., news vertical). In other words, the relevance of verticals differ with queries. The main challenge in aggregated search is how to identify and integrate relevant heterogeneous results for each given query into a single result page.

Aggregated search has three main components: (1) *vertical representation*, concerned with how to represent verticals so that the documents contained within and their type are identifiable, (2) *vertical selection*, concerned with how to select the verticals from which relevant documents can be retrieved, and (3) *result presentation*, concerned with how to assemble results from selected verticals so as to best layout the result page with the most relevant information. These components are described in Sections 3, 4 and 5, respectively. We also provide some background about aggregated search and related paradigms in Section 2, and discuss evaluation of aggregated search in Section 6. This chapter finishes with some conclusions.

2 Background and motivation

Aggregated search seeks relevant content across heterogenous information sources, the verticals. Searching diverse information sources is not new. Federated search (also referred to as distributed information retrieval [10]) and metasearch are techniques that aim to search and provide results from various sources.

In federated search, a user submits a query, and then may select a number of sources, referred to as resources, to search. These resources are often standalone systems (e.g. corporate intranets, fee-based databases, library catalogues, internet resources, user-specific digital storage). The federated search system, when not explicitly stated by the user, has to identify the most relevant resources (those with the highest number of relevant documents) to search given a query (resource selection). It then sends the query to those (one or several) selected resources. These resources return results for that query to the federated search system, which then decides which and how many results to retain. These selected results are presented to the user. The results are often returned merged within one single ranked list, but can also be separated, for example, grouped per resource where they originate. In some cases, resources may return duplicate results, which should be removed. Examples of federated search systems include Funnelback[2], Westlaw[3], FedStats[4]. We refer the reader to [38]

[2] http://www.funnelback.com/
[3] http://www.westlaw.co.uk/
[4] http://www.fedstats.gov/

for an extensive survey on federated search and the Federated Search Blog at http://federatedsearchblog.com/ for latest developments in the federated search industry.

Bringing federated search to the web led to two different paradigms, metasearch and aggregated search [38]. A metasearch engine is a search engine that queries several different search engines, and combine results from them or display them separately. A metasearch engine operates on the premise that the web is too large for any one search engine to index, and that more comprehensive search results can be obtained by combining results from several search engines. This also saves the user from having to use multiple search engines separately. Metasearch engines were more popular 10-15 years ago as now the partial coverage of the web space seems less of an issue with current major search engines (Google, Yahoo!, Bing[5]) compared to earlier ones (Altavista, Lycos, etc). In addition, unless some agreements are in place, current search engines usually do not provide unlimited access of their search results to third party applications, such as a metasearch engine, because of incurred traffic loads and business models. Examples of existing metasearch engines include Dogpile[6], Metacrawler[7], and Search.Com[8].

An aggregated search system also provides information from different sources. However, in aggregated search, the information sources are powered by dedicated vertical search engines, all mostly within the remit of the general web search engine, and not several and independent search engines, as is the case with metasearch. In addition, the individual information sources in aggregated search retrieve from very different collections of documents, e.g. images, videos, news. A typical example of an aggregated search system is shown in Figure 1. Here is the result page for the query "world cup" issued just after the final world cup football game in July 2010 to Google. We can see a mixture of structured data (editorial content), news, homepage, wikipedia, real-time results, videos and tweets.

The idea of aggregated search was explicitly introduced as universal search in 2007 by Google[9]:

> "[...] search across all its content sources, compare and rank all the information in real time, and deliver a single, integrated set of search results [...] will incorporate information from a variety of previously separate sources including videos, images, news, maps, books, and websites – into a single set of results."

The goal behind aggregated search is to remedy the fact, that overall, vertical search is not prevalently used by users. Indeed, JupiterResearch [5] carried out a survey in 2007-2008 that indicates that 35% of users do not use vertical

[5] Not that since August 2010, Yahoo! search engine is powered by Bing.
[6] http://www.dogpile.com/
[7] http://www.metacrawler.com/
[8] http://www.search.com/
[9] http://www.google.com/intl/en/press/pressrel/universalsearch_20070516.html

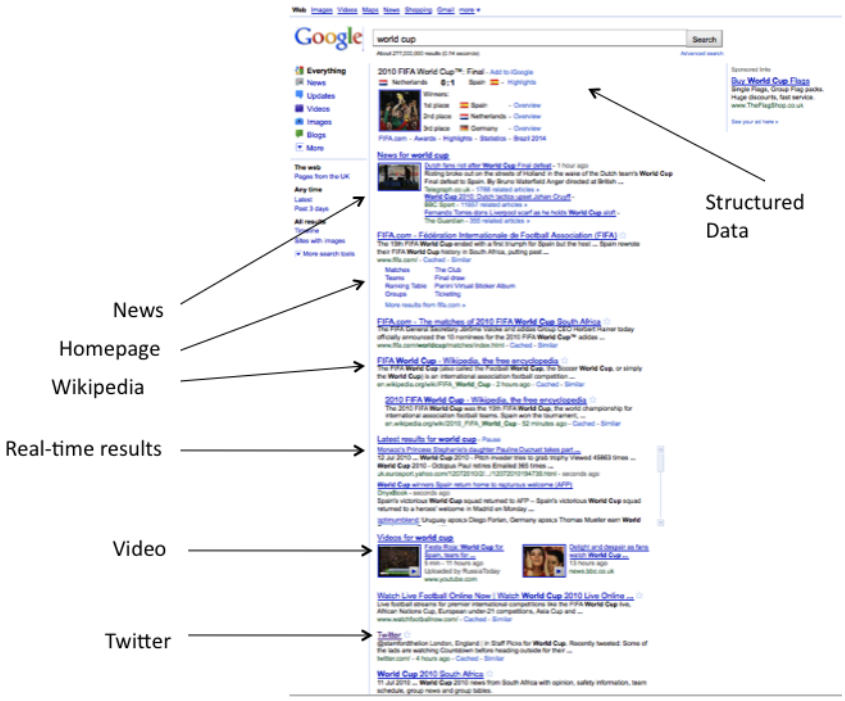

Fig. 1. Example of aggregated search (blended design) - Google

search. This does not mean that users do not have queries with one or more vertical intents. The fact that queries can be answered from various verticals was shown, for instance, in [8], who looked at 25,195 unique queries obtained from a commercial search engine query log. Human editors were instructed to assign between zero and six relevant verticals per query based on their best guess of the user vertical intent. About 26% of queries, mostly navigational, were assigned no relevant vertical and 44% were assigned a single relevant vertical. The rest of the queries were assigned multiple relevant verticals, as they were ambiguous in terms of vertical intent (e.g., the query "hairspray" was assigned the verticals movies, video, and shopping [8]). Query logs from a different commercial search engine analysed in [32] showed that 12.3% of the queries have an image search intent, 8.5% have a video search intent and so on (the total number of analysed queries was 2,153). Similar observations were reached in [44], who analysed query log click-through in terms of vertical intents. Thus, a vertical search intent is often present within web search queries.

Within one year of major search engines providing users with aggregated search results, a greater percentage of users clicked on vertical search result types within the general search results, compared to when the verticals were searched directly [5]. For example, news results were the most clicked vertical results within aggregated search, and users click them more than twice as much within

aggregated search than they do when they use the vertical news search directly. More recent statistics showing similar trends can be found in a ComScore report [23]. Thus, despite users limited use of vertical search engines, it is important for search engines to ensure that their relevant vertical contents are being shown.

3 Vertical representation

To select the relevant vertical(s) for each submitted query (described in Section 4), an aggregated search engine needs to know about the content of each vertical (e.g. term statistics, size, etc). This is to ensure that, for example, the query "tennis" is passed to a sport vertical, whereas the query "Madonna" is sent to a music vertical and eventually a celebrity vertical. For this purpose, the aggregated search system keeps a representation of each of its verticals.

Vertical representation can be compared to resource representation in federated search, in which a number of techniques have been proposed [38]. In federated search, a resource representation can be generated manually by providing a short description of the documents contained in that resource (e.g. [24]). However, these representations are likely to be brief, thus providing only a limited coverage of the documents contained in the resource, and more importantly, quickly become stale for dynamic resources, where new documents are often added, and in large number. In practice, automatically generated representations are more common.

In aggregated search, various statistics about the documents (e.g. term frequency, vertical size) contained in the vertical are available – for those verticals operated by the same body – and are used to generate the vertical representation. Therefore, a vertical representation can be built using techniques from federated search working with cooperative resources[10]. A technique reported in the literature is the generation of vertical representations from a subset of documents, so-called sampled documents. Two main sampling approaches have been reported in the literature [8], one where the documents are sampled directly from the vertical, and another using external resources.

To directly sample from a vertical, query-based sampling is used. In the context of federated search [11], this works as follows. An initial query is used to retrieve documents from the resource, which are then used to generate an initial representation of the resource (or update the current representation if the aim is to refresh the representation to account for newly added documents). Then, a so-called new sampling query is selected from the representation. As documents are retrieved for each sampling query, the evolving resource representation and sampling queries are derived from the retrieved documents. It was, however, shown [39] that better performance were obtained when high-frequency queries were used for sampling the documents, than when derived from the sampled documents themselves. Similarly, in the context of aggregated search, high-frequency

[10] The resources provide to the federated search system comprehensive statistics about their documents e.g. term frequency, size.

queries, extracted from vertical query logs, have been used for sampling documents [8]. Indeed, queries issued directly to a vertical represent explicit vertical intent, and therefore constitute good candidates to sample documents for that vertical.

Although initially introduced to deal with un-cooperative resources, and even though that (most) verticals can be viewed as cooperative resources, the sampling approach is particularly appropriate for aggregated search. Indeed, using high-frequency vertical queries leads to sampled documents that are biased towards those that are more likely to be useful to users, and thus more likely to be seen by these users [8]. This is important because it is likely that a significant part of the vertical is of no interest to users, and thus should not be represented. In addition, using high-frequency vertical queries on a regular basis (how regularly depends on the vertical), is a way to ensure that vertical representations are up-to-date, in particular for verticals with a highly dynamic content, such as news. For this type of verticals, users are mostly interested in the most recent content.

An alternative, or in addition to sampling directly from the verticals, is to sample documents from external sources, if documents can be mapped to verticals. This approach was investigated in [8] who sampled documents from Wikipedia, making use of Wikipedia categories (one or several) assigned to documents to map documents to verticals. For instance, a sample of documents representative of the "car" vertical can be gathered from documents assigned a Wikipedia category containing any of the terms "auto", "automobile", "car", and "vehicle". Wikipedia is rich in text, so sampling from it can help in providing better representations of text-impoverished verticals such as image and video verticals. In addition, Wikipedia articles have a consistent format, are semantically coherent and on topic. This means that a better coverage of the vertical content can be achieved, as well as a more uniform representation across verticals. The latter can also make comparing rankings across verticals easier.

Overall, using high-frequency vertical queries to sample the vertical directly, together with sampling from external sources such as Wikipedia, can ensure that the most relevant and recent vertical content can be retrieved (popular and/or peak queries), while still providing a good coverage of the vertical content (useful for tail queries), and that text-impoverished verticals can be properly represented (image and video verticals).

4 Vertical selection

Vertical selection is the task of selecting the relevant verticals (none, one or several) in response to a user query. Vertical selection is related to resource selection in federated search, where the most common approach is to rank resources based on the similarity between the query and the resource representation. Some approaches [12, 41] treat resources, i.e. their representations, as "large documents" and adapt ranking functions from IR to score resources. More effective tech-

niques, such as ReDDE [40], consider the distribution of relevant documents across resources to score them.

Vertical selection also makes use of the vertical representations, in a way similar to federated search, but has access to additional sources of evidence. Indeed, verticals focus on specific types of documents (in terms of domain, media, or genre). Users searching for a particular type of document (e.g. "images of red cars") may issue domain/genre/media specific words in the query (e.g. "picture"). The query string therefore constitutes a source of evidence for vertical selection. Second, vertical search engines are being used explicitly by users, so associated query logs may be available, which can be exploited for vertical selection. These two sources of evidence can be used to provide directly to the aggregated search engine a so-called "vertical relevance value", by the vertical search engine, reflecting how relevant the vertical content might be to the query.

How these two sources of evidence, and the vertical representations, are used to select verticals has been reported in a series of papers [8, 20, 9]. Machine learning techniques are used to build models of verticals based on features extracted from vertical representations, query strings and vertical query logs. More precisely [8], features from the vertical representations include vertical ranking scores such as those produced by ReDDE in federated search [40]. Features for the query string features were based on rules triggered by word occurrences in the query. These rules mapped words to verticals (e.g. "car" to the autos vertical). Geographical words were also used; for example, queries with the words "continent", "country", "county", "point of interest" were mapped to verticals related to local, travel and maps. Finally, features for the vertical query logs correspond to the query likelihood built from a unigram language model constructed from the vertical query logs.

The findings showed that ranking verticals by the query likelihood estimated from the vertical query language model was the best single-evidence to select a vertical. It was also shown that using rules mapping query strings to verticals led to significant improvement in vertical selection. This is particularly useful in situations where no training data is available for a vertical. With respect to the latter, research reported in [9] looked at how models learned for a vertical with training data could be "ported" to other verticals for which there is no training data. The technique of machine adaptation, from machine learning, was employed, which consists of using training data from one or more source domains to learn a predictive model for a target domain, typically associated with little or no training data.

5 Result presentation

For a large percentage of queries, relevant results mostly come from the conventional web. However, for an increasing number of queries, results from other sources (verticals) are relevant, and thus should be added to the standard web results. Research reported in [19] showed the positive effect (in relevance ranking) of properly integrating news within web results, and in [42] demonstrated

through a user study the effectiveness of aggregated search for non-navigational queries. Thus, in aggregated search, a main challenge is to identify the best positions to place items retrieved from relevant verticals on the result page to e.g. maximising click-through rate.

There are two main types of result page design in aggregated search, one where results from the different verticals are blended into a single list, referred to as *blended*, and another, where results from each vertical are presented in a separate panel, referred to as *non-blended*.

A blended integration as applied by Google universal search and many other search engines presents results from the different verticals within a single ranked list. It should be pointed out that blending is not the same as inter-leaving. In the blended design of aggregated search, results from the same vertical are usually "slotted" together in the ranking, as can been seen in Figure 1. The blended design seems the most common way to present results in aggregated search.

With the blended design, relevance remains the main ranking criteria within verticals and across verticals. Results from the same vertical are slotted together (e.g. 4 news, 6 images, etc, each ranked with respect to the their estimated relevance to the query), but the entire slot is ranked with respect to other slots, including standard web search results. Other ranking criteria may be used, for example, newsworthiness for results coming from a news vertical. For example, in our example in Figure 1, the newsworthiness of the query "world cup" just after the final game in July 2010 was very high, and as such, news were placed towards the top of the result page (in addition to editorial content). Today, submitting the same query to Google returns a completely different result page (the top result is the official FIFA site).

Two approaches for blended aggregated search have been reported in the literature. In [33], using machine learning techniques, probabilities such as a document in a vertical being relevant, a vertical being relevant, etc., to a query, were estimated, and used in a probabilistic model. The resulting probabilistic document scores were used to rank all documents, from the standard web and vertical results, into one single list. Although an increase in performance is observed, the quality of the blended search results is not really evaluated. Indeed, it is not clear whether the returned results are optimal in terms of aggregated search (i.e. the right combination of web search results and vertical results, and at the right position). It should also be added that results were inter-leaved (as given by the calculated probabilistic scores), and not blended as above described.

A recent work [35] investigates blended aggregated search, where the relationship between results from the web and verticals is accounted for. The focus of their work is that given multiple already known to be relevant verticals, how to place results from them relative to web results. Machine learning techniques are used on training data based on elicited pairwise preferences between groups of results, i.e. a group of standard web results and a group of vertical results. When ranking, what is being compared is a group of web results and a group of vertical results. For the final composition of the blended result page, a group of vertical results is placed at a slot if the score is higher than some threshold

employed to guarantee specific coverage for verticals at each slot. Features used include query log-based features, vertical-based features (provided by the vertical partner), and query string-based features. They show that using pairwise preferences judgements for training significantly improve retrieval performance, and increase user interaction with results. This the first work publishing detailed account on how blended result pages are composed.

The non-blended integration presents results from each vertical in a separate panel. In search engine terminology, the panel is also referred to as a "tile". Alpha Yahoo![11] shown in Figure 2 is an example of such a design. Other examples include Kosmix[12], Naver[13] and the discontinued Google Searchmash[14]. Whenever a minimal amount of results from a vertical is available for a query, the various corresponding panels are filled and displayed. The main web search results are usually displayed on the left side and within the largest panel, conveying to users that most results still come from the general web. There is no relationships between results from the different panels. The placement of the various panels is also mostly predefined.

Although a large number of studies devoted to the design and evaluation of conventional web search interfaces have been reported in the literature, less is known about aggregated search interfaces, apart for maybe three studies, two eye-tracking and one within-subject task-based experiments.

An eye-tracking experiment on Google Universal search soon after its launch has been reported [27]. Screenshots of the eye-tracking outcomes (users visual attention on the result page) are shown in Figures 3 and 4, where the main difference is the addition of image results (thumbnails) in the result page. In the pre-aggregated search interface (right screenshot of Figure 3), the common trend is to start in the upper left corner (indicated by A) and to scan results from there, first vertically (the down arrow) and then across – likely when a title catches the user attention. A distinct pattern is observed with the aggregated search interface (left screenshot of Figure 3). While there is still some scanning in the very upper left (B), the scanning does not start there, but from around the image results (C). Scanning seems to be predominantly to the side and below of the thumbnail (D). Furthermore, the F pattern [34] for scanning results in conventional web interface seems to change to an E pattern in an aggregated interface (Figure 4).

However, another eye-tracking study reported by Google [1] observed that, returning results from verticals blended within typical web search results did not affect the order of scanning the results, neither did it disrupt the information seeking process of users. This study was carried in 2009, where users by then, would have become more familiar with being returned results from verticals.

These studies provide insight into how users view results in an aggregated search interface. There is some understanding on where on the result page the

[11] http://au.alpha.yahoo.com/
[12] http://www.kosmix.com/
[13] http://www.naver.com/
[14] http://techcrunch.com/2008/11/24/why-did-google-discontinue-searchmash/

Fig. 2. Example of aggregated search (non-blended design) - Yahoo! Alpha

user looks at or gives attention to, and where the user starts viewing results from verticals. However, whether and how results are presented affects user behaviour (e.g. click-through rate, task completion, etc) was not discussed in these studies, and to our knowledge, nothing has been reported about this in the literature. Therefore, to provide some insight into the latter, we carried out two within-subject task-based user studies [43], one using a blended design and another using a non-blended design. Our objective was to investigate the impact of factors on user click-through behaviour on these two types of aggregated search interface. The factors studied included position of search results, vertical types, and the strength of search task orientation towards a particular vertical.

Studies, e.g. log analysis and eye-tracking experiments, which look at the effect of result position in the context of conventional web search, are not new. For instance, [25] showed that when results were placed relatively low in the result page, people spent more time searching and were less successful in their search task. Similar behaviour is likely to be observed with aggregated search interfaces. This has motivated us to investigate the position effect across results from different verticals.

Although, compared to in-house investigations carried out by commercial search engine companies, our experiment is small (a total of 1,296 search sessions

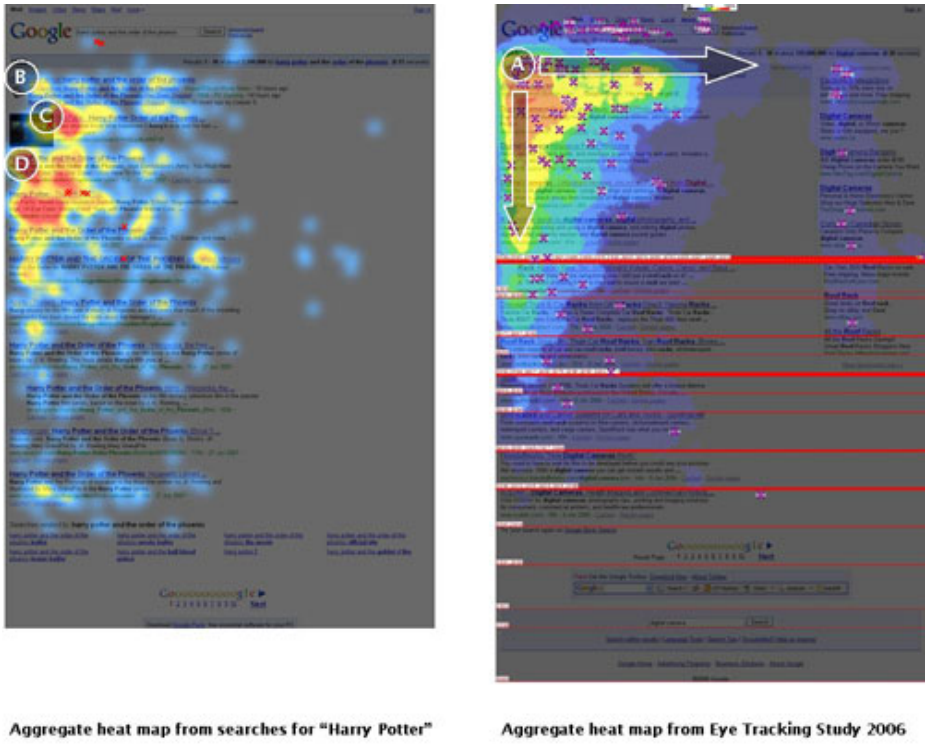

Fig. 3. Eye-tracking study comparing Google universal search to "standard" search – Taken from [27]

performed by 48 participants were analysed, and 3 verticals were considered, image, video and news), we obtained interesting findings.

Our first finding is that the factors that affect user click-through behavior differ between the blended and non-blended designs. Behaviours observed with the blended design echoed the findings of previous studies in standard web search interfaces e.g. [28, 6, 25, 30], but not in the non-blended design. This suggests that a careful estimation of the relevance of verticals is needed when the blended design is employed. When we cannot measure their relevance, the non-blended design is more appropriate since users click-through behaviour was not affected by the position in this type of aggregation.

The second finding is that videos resulted in a different click-through pattern from news and images. This trend was common in both the blended and non-blended designs. This suggests that, when deciding to retrieve videos, different behaviour from other verticals may be observed. A different pattern with video results was also observed in [23], compared to standard web results. In their case, video results with thumbnails generated higher click-through rates, across

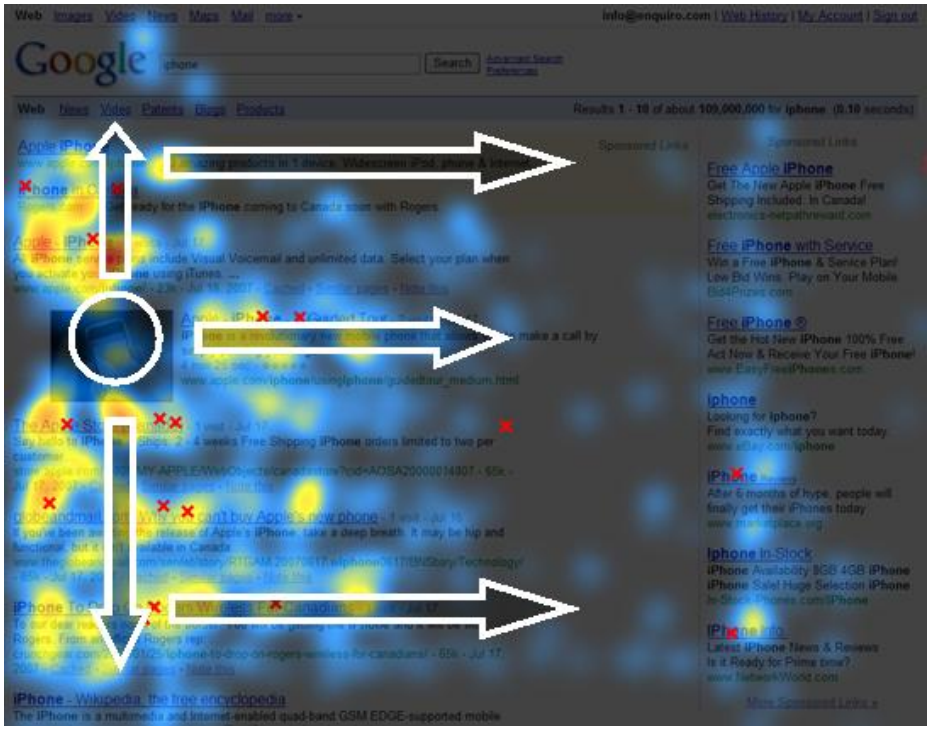

Fig. 4. Google universal search eye-tracking study: E scanning pattern – Taken from [27]

all positions, and especially helped results in lower positions. More generally, the above suggests that click-through behaviour can be different across verticals, and this should be accounted for by aggregated search systems.

The third finding was that a search task orientation towards a particular vertical can affect click-through behaviour. This trend was common to both the blended and non-blended designs. Traditional information retrieval research has been focused on the modelling of thematic (or topical) relevance of documents. However, research in other areas has demonstrated that relevance can be multi-dimensional [18], e.g. in XML retrieval [31] and geographic information retrieval [36]. In a similar way, experiments on aggregated search should control the level of orientation towards a particular vertical, as this is an important factor to investigate, not only with respect to algorithms (e.g. vertical selection), but also result presentation. Ultimately, developing interaction design that can help aggregated search systems capture a user vertical preference of an information need would be welcome.

6 Evaluation

IR has a long evaluation history [17]. Advances in IR research (in terms of effectiveness) have mostly been made through experimental evaluation on test collections. A test collection is created to allow experimentation with respect to a retrieval scenario, e.g. blog retrieval, XML retrieval. Given that scenario, a collection of items that are representative for this scenario is collected. Then, a set of topics is developed, reflecting information needs typical of the scenario. These topics are created by humans, who, when possible, are domain experts. Finally, using a pooling technique, a subset of items (the pool containing the items that are most likely to be relevant) is judged by humans to whom various incentives are given (e.g. monetary or access to data), and who, when possible, are the creators of the topics. The TREC evaluation initiative from NIST [46] has been a major player in IR evaluation, and its methodology has been adopted by many other initiatives, such as CLEF [3], INEX [31], and NCTIR [4], to name a few.

The most time-consuming part of generating a test collection is the creation of relevance judgments. Indeed, for every topic, a large number of items (even though a pool is used) must be judged. Although methods to alleviate this problems have been proposed (e.g. formally selecting a subset of the most promising items to be judged [13] or using crowd-sourcing techniques [7]) judging a set of items (e.g. documents), and in particular, heterogeneous documents from a variety of sources, remains an extremely tedious task. Constructing such a test collection from scratch is very time-consuming. For example, INEX ran a heterogenous track [22], that never really flourished due to the required effort of having to create topics with relevant documents across multiple heterogeneous XML repositories and the laborious task of providing relevance assessments for the topics.

A test collection for aggregated search requires verticals, each populated by items of that vertical type, a set of topics expressing information needs relating to one or more verticals, and relevance assessments, indicating the relevance of the items and their associated verticals to each topic. Although work on aggregated search has been conducted and experimental results reported (e.g. vertical selection [8], result presentation [35]), no large-scale test collection for aggregated search is available.

Evaluation of aggregated search in terms of effectiveness so far reported in the literature has focused on vertical selection accuracy [8], but also on the quality of the blended search results [33, 35]. The data sets, aka the test collections, used for the evaluation are those typical to commercial search engines. They include a mixture of editorial data (e.g. labelers judged the relevance between queries and verticals, or pairwise preferences between groups of results), and behavioural data (e.g. query logs inferred from click data). Nonetheless these data sets are not widely accessible.

Following the spirit of TREC and similar initiatives, we recently proposed a methodology to build a test collection for aggregated search [47]. Our methodology makes use of existing test collections available from TREC, CLEF and

INEX, where a vertical is simulated by one, part of, or several existing test collections. Reusing existing test collections to build new ones is not new. Indeed, because current test collections (e.g. Clueweb09 [2]) have become extremely large to reflect the much larger amount of information in many of today's retrieval scenarios, the idea of reusing test collections is very appealing. For example, [16] reused an existing Q&A test collection to generate a test collection to investigate diversity in IR, and [14] developed means to quantify the reusability of a test collection for evaluating a different retrieval scenario than that originally built for. In federated search, test collections (e.g. trec4-kmeans [40]) have also been developed using existing test collections by partitioning different text-based TREC corpora into a number of sub-collections. However, these test collections cannot be applied to aggregated search as all sub-collections are of the same type. The main difference between these types of (federated vs. aggregated) search is the heterogeneous natures of the items, as each vertical comprises of items of a particular genre, domain, and/or media.

As stated earlier there are three parts to a test collection, the collection of items, the topics and the relevance assessments. Our first step was to decide which existing (or part) test collections could be used to simulate verticals. We used, for example, ImageCLEF to simulate an image vertical, INEX Wikipedia to simulate a wiki vertical, TREC blog to simulate a blog vertical, etc. We also used Clueweb09; however as this collection is made of text-based documents of different genres, we employed a classifier to assign documents to verticals related to genre, e.g. news, references, wikipedia. In total, we obtained a total of 80 million documents, with ten simulated verticals. General web documents were prevalent, thus mimicking aggregated search scenarios.

We selected topics (from all available topics coming from the used test collections) that we believed would reflect concrete search scenarios in aggregated search, i.e. topics for which only "general web" documents are relevant, and those for which documents from more than one vertical are relevant. For the former, we randomly chose topics from Clueweb09, submitted their titles to a search engine, and retained only those that returned only standard web results. For the latter, various strategies were applied, depending on wether topics "co-exist" in several test collections, or topics for which relevant documents of different genres exist (e.g. all of which come from Clueweb09). At the end, 150 topics were collected, 67% with two vertical intents (i.e. contain relevant documents in two verticals) and 4% with three or more vertical intents. These statistics compare to those in [8], coming from editorial data.

The relevance assessments were simply those coming from the chosen collections and selected topics. One main drawback with this approach, however, is that of incomplete relevance assessments with respect to a whole (simulated) vertical. Indeed, it can happen that a topic coming from test collection A is not a topic of a second test collection B, but B may actually contain documents relevant to that topic, where A and B have been used to simulate two different verticals. Initial experiments showed that this did not have too strong adversarial affect, but providing additional relevance assessments may be required.

Nonetheless, performing these additional assessments remains less costly compared to building a test collection for aggregated search from scratch.

When evaluating retrieval effectiveness using a test collection, metrics are employed, e.g. precision/recall and variants of them. In the standard scenario, a ranked list of results is returned for each topic forming the test collection used for the evaluation. These rankings are individually evaluated using the metrics, and some average effectiveness value across topics is calculated afterwards (e.g. MAP). The returned items are usually considered separately, but not always (e.g. in XML retrieval [29], and the earlier TREC web track [26]). In aggregated search, simply looking at the results individually is definitively not sufficient. For example, relations between the retrieved items may be important to deliver better rankings (as shown in [35]); this goes beyond relevance. Indeed, sometimes some items have to be shown before others even if less relevant (e.g. because of chronological order), or should be grouped together (similar content, similar type of content, alternative, A is a picture of B, etc). Devising metrics that consider these relationships when evaluating retrieval effectiveness is challenging.

Finally, there has been a recent increased interest toward the notion of result diversification not only in web search e.g. [15], but also e.g. geographical [45] and image search [37]. Diversification is believed to maximise the likelihood for ambiguous queries returning at least one relevant result, while also showing the multiple aspects of queries. Although web search queries may often hide vertical intents, the impact of diversity, in the context of aggregated search, i.e. returning results from several verticals (vertical diversity) remains to be explicitly assessed.

7 Conclusions

For a large percentage of queries, relevant results mostly come from the conventional web. However, for an increasing number of queries, results from other sources, the so-called verticals, are relevant, and thus should be added to the standard web results. Aggregated search aims to facilitate the access to the increasingly diverse content available from the verticals. It does so by searching and assembling relevant documents from a variety of verticals and placing them into a single result page, together with standard web search results. The goal is to best layout the result page with the most relevant information from the conventional web and verticals. Most current search engines perform some level of aggregated search. We have surveyed the state of the art in aggregated search. We expect this to be a rich and fertile research area for many years to come.

Acknowledgements: This chapter was written when Mounia Lalmas was a Microsoft Research/RAEng Research Professor of Information Retrieval at the University of Glasgow. Her work on aggregated search was carried out in the context of a project partly funded by a Yahoo! Research Alliance Gift. This chapter is inspired by a tutorial "From federated search to aggregated search" presented at the 33rd ACM SIGIR Conference on Research and Development in Information Retrieval, 19-23 July

2010, Geneva, Switzerland [21]. Special thanks go to Ke (Adam) Zhou and Ronnan Cummins for their feedbacks on earlier versions of this paper.

References

1. A. Aula and K. Rodden. Eye-tracking studies: more than meets the eye. http://googleblog.blogspot.com/2009/02/eye-tracking-studies-more-than-meets.html, June 2009.
2. J. Callan. The ClueWeb09 Dataset. http://boston.lti.cs.cmu.edu/Data/clueweb09/, 2010.
3. Cross-Language Evaluation Forum (CLEF). http://www.clef-campaign.org/
4. http://research.nii.ac.jp/ntcir/index-en.html.
5. iProspect Blended Search Results Study. http://www.iprospect.com/about/researchstudy_2008_blendedsearchresults.htm, April 2008.
6. E. Agichtein and Z. Zheng. Identifying best bet web search results by mining past user behavior. SIGKDD, pp 902–908, 2006.
7. O. Alonso, D. Rose and B. Stewart. Crowdsourcing for Relevance Evaluation. SIGIR Forum 42(2): 9–15, 2008.
8. J. Arguello, F. Diaz, J. Callan and J. Crespo. Sources of Evidence for Vertical Selection. SIGIR, pp 315–322, 2009.
9. J. Arguello, F. Diaz and J. F. Paiement. Vertical selection in presence of unlabeled verticals. SIGIR, pp 691–698, 2010.
10. J. Callan Distributed Information Retrieval. *Advances in Information Retrieval*, pp 127–150, 2000
11. J. Callan and M. Connell. Query-based sampling of text databases. TOIS 19(2):97–130, 2001.
12. J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. SIGIR, pp 21–28, 1995.
13. B. Carterette, J. Allan and R. Sitaraman. Minimal Test Collections for Retrieval Evaluation. SIGIR, pp 268–275, 2006.
14. B. Carterette, E. Gabrilovich and D. Metzler. Measuring the Reusability of Test Collections. WSDM, pp 231–240, 2010.
15. C.L.A. Clarke, N. Craswell and I. Soboroff. Overview of the TREC 2009 Web Track. TREC, 2009.
16. C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher and I. MacKinnon. Novelty and Diversity in Information Retrieval Evaluation. SIGIR, pp 659–666, 2008.
17. C. Cleverdon. The Significance of the Cranfield Tests on Index Languages. SIGIR, pp 3–12, 1991.
18. E. Cosijn and P. Ingwersen Dimensions of relevance. IP&M, 36:533–550, 2000.
19. F. Diaz. Integration of news content into web results. WSDM, pp 182–191, 2009.
20. F. Diaz and J. Arguello. Adaptation of offline selection predictions in presence of user feedback. SIGIR, pp 323–330, 2009.
21. F. Diaz, M. Lalmas and M. Shokouhi. From federated to aggregated search. SIGIR, pp 910, 2010.
22. I. Frommholz and R.R. Larson Report on the INEX 2006 heterogeneous collection track. SIGIR Forum, 41(1):75-78, 2007.

23. E. Goodman and E. Feldblum. Blended Search and the New Rules of Engagement. ComScore Report, October 2010.
24. L. Gravano, C. Chang, H. Garca-Molina and A. Paepcke. STARTS: Stanford proposal for internet metasearching. SIGMOD, pp 207–218, T1997
25. Z. Guan and E. Cutrell. An eye tracking study of the effect of target rank on web search. SIGCHI, pp 417–420, 2007.
26. D. Hawking, E.M. Voorhees, N. Craswell and P. Bailey. Overview of the TREC-8 Web Track. TREC, 1999.
27. G. Hotchkiss. Eye Tracking On Universal And Personalized Search. Search Engine Land. http://searchengineland.com/eye-tracking-on-universal-and-personalized-search-12233, September 2007.
28. T. Joachims, L. Granka, B. Pan, H. Hembrooke and G. Gay. Accurately interpreting clickthrough data as implicit feedback. SIGIR, pp 154–161, 2005.
29. G. Kazai, M. Lalmas and A.P. de Vries. The overlap problem in content-oriented XML retrieval evaluation. SIGIR, pp 72-79, 2004.
30. M.T. Keane, M. OBrien and B. Smyth. Are people biased in their use of search engines? CACM 51(2):4952, 2008.
31. M. Lalmas and A. Tombros. INEX 2002-2006: Understanding XML retrieval evaluation. International conference on Digital libraries, pp 187–196, 2007.
32. N. Liu, J. Yan, W. Fan, Q. Yang, and Z. Chen. Identifying Vertical Search Intention of Query through Social Tagging Propagation. WWW, 2009.
33. N. Liu, J. Yan, Z. Chen A probabilistic model based approach for blended search. WWW, pp 1075-1076, 2009.
34. Jakob Nielsen. *Eyetracking Web Usability*. Kara Pernice, 2009.
35. A.K. Ponnuswami, K. Pattabiraman, Q. Wu, R. Gilad-Bachrach and T. Kanungo. On Composition of a Federated Web Search Result Page: Using Online Users to Provide Pairwise Preference for Heterogeneous Verticals. WSDM, 2011
36. R. Purves and J. Chris Workshop on geographic information retrieval, SIGIR 2004 SIGIR Forum 38(2): 53–56, 2004.
37. M. Sanderson, J. Tang, T. Arni and P. Clough. What Else Is There? Search Diversity Examined. ECIR, pp 562-569, 2009.
38. M. Shokouhi and L. Si. *Federated Information Retrieval*. Foundations and Trends in Information Retrieval, Upcoming Issue, 2011.
39. M. Shokouhi, J. Zobel, S. Tahaghoghi and F. Scholer. Using query logs to establish vocabularies in distributed information retrieval. IP&M, 43(1):169180, 2007.
40. L. Si and J. Callan. Relevant Document Distribution Estimation Method for Resource Selection. SIGIR, pp 298–305, 2003.
41. L. Si, R. Jin, J. Callan, and P. Ogilvie. A language modeling framework for resource selection and results merging. CIKM, pp 391–397, 2002.
42. S. Sushmita, H. Joho and M. Lalmas A Task-Based Evaluation of an Aggregated Search Interface. SPIRE, pp 322-333, 2009.
43. S. Sushmita, H. Joho, M. Lalmas and R. Villa. Factors Affecting Click-Through Behavior in Aggregated Search Interfaces CIKM, pp 519-528, 2010.
44. S. Sushmita, B. Piwowarski and M. Lalmas. Dynamics of Genre and Domain Intents, AIRS, 2010.
45. J. Tang and M. Sanderson. Evaluation and User Preference Study on Spatial Diversity. ECIR, pp 179-190, 2010.
46. E. Voorhees and D. Harman. *TREC: Experiments and Evaluation in Information Retrieval*. MIT Press, 2005.
47. K. Zhou, M. Lalmas and R.Cummins. Building a Test Collection for Aggregated Search. Research Report, University of Glasgow, October 2010.