Interest and Evaluation of Aggregated Search

Arlind Kopliku, Firas Damak*, Karen Pinel-Sauvagnat**, Mohand Boughanem***
IRIT
University of Toulouse
Toulouse, France
*Arlind.Kopliku@irit.fr, * Firas.Damak@irit.fr, ** Karen.Sauvagnat@irit.fr, *** Mohand.Boughanem.irit.fr*

***Abstract*—Major search engines perform what is known as Aggregated Search (AS). They integrate results coming from different vertical search engines (images, videos, news, etc.) with typical Web search results. Aggregated search is relatively new and its advantages need to be evaluated. Some existing works have already tried to evaluate the interest (usefulness) of aggregated search as well as the effectiveness of the existing approaches. However, most of evaluation methodologies were based (i) on what we call relevance by intent (i.e. search results were not shown to real users), and (ii) short text queries.**

In this paper, we conducted a user study which was designed to revisit and compare the interest of aggregated search, by exploiting both relevance by intent and content, and using both short text and fixed need queries. This user study allowed us to analyze the distribution of relevant results across different verticals, and to show that AS helps to identify complementary relevant sources for the same information need. Comparison between relevance by intent and relevance by content showed that relevance by intent introduces a bias in evaluation. Discussion about the results also allowed us to identify some useful thoughts concerning the evaluation of AS approaches.

***Keywords*-aggregated search, vertical selection, evaluation, user study**

I. Introduction

The success of search engines was initially due to Web search, but recently many vertical searches are becoming popular. They focus search on a specific topic (legal information, medical information, ...), or type of media (video, images, news, blogs, ...), or geographical location, industry and so on. In constrast to general (horizontal) web search that aims at crawling and indexing the whole WWW, vertical searches attempt to crawl and index only web pages related to their predefined domains. Most of these vertical searches (maps, blogs, images, videos, ...) can now be easily accessed from major search engines.

However, to improve search, there is a trend today to directly integrate vertical search results within Web search results [13], [2], [14]. This is called Aggregated Search (AS). This enables users to query several sources (vertical search engines) from one interface. One of the questions of AS is to predict which sources are likely to provide relevant items to queries. For instance, the query "Eiffel tower photos" can be answered from image search. Some other queries can be answered from multiple sources; *"visit France"* for instance can be answered with Web pages as well as with images, videos or maps. Research in aggregated search has taken three main directions. One direction studies adequate interfaces for aggregated search results visualization [14], [16]. Another direction involves techniques that can predict whether results from one source (vertical) should be included or not in response to a given query [2], [11], [8]. This can also be viewed as a form of source selection in distributed information retrieval [4]. The last direction concerns the evaluation of the interest (usefulness) of AS as well as the definition of methodologies to evaluate AS approaches.

The research question investigated in this paper falls in the latter direction. The interest of AS was studied in [2], [11], [12]. Thousands of queries issued from query logs were assessed. Each query was assigned one or more vertical intent (a vertical search that is likely to answer the query). Results show that vertical search intent is often present in Web search queries. Although participants assessed many queries, they did not know the need behind the query and they were not shown any concrete results from search engines. Relevance was only evaluated *by intent*.

In [14], [16], Sushmita et al. investigated some advantages of AS interfaces. They showed that AS increases the quantity and diversity of relevant results accessed by users. Here relevance is evaluated *by content* (i.e. by considering the content of results) and queries are associated with an information need.

Our goal in this paper is to reconsider the evaluation of the interest of AS, by exploiting both relevance by intent and by content, and by using queries with or without fixed need. Our research questions include:

- how are relevant results distributed across different sources ?
- which is the interest of using different vertical searches compared to traditional web retrieval ?
- Are different sources complementary to each other? In other words, is one source sufficient to answer a given query or does it need to be completed with other sources ?
- How do relevance assessments and query types impact the evaluation of aggregated search?

Some of those questions have already been examined in literature (see section II-C for details). Our aim is to revisit them by exploiting two definitions of relevance and two types of queries. For this purpose we conducted a user study, which involved human participants. We examined four search situations, by considering the two types of queries (short text and fixed need) and the two types of relevance (by intent and content).
This paper is structured as follows. We initially introduce related work concerning AS. Sections III and IV introduce our user study and its results. In section V we discuss results and give some thoughts about the evaluation of AS.

II. Related work

As mentioned in the introduction, aggregated search leads to at least three research questions: how to predict sources that are likely to provide relevant items ? how to visualize the results ? and how to evaluate approaches ?

In the following we will briefly discuss related works for the first two questions, and we will then focus on evaluation.

A. Source (vertical) selection

One of the main issues in aggregated search is to select the relevant sources for an information need. This corresponds to *vertical selection* [2]. We prefer using the term *source selection* as it allows to generalize over both verticals and Web search.
Source selection consists in predicting if a source is relevant or not for a given query.
The prediction of the correct source is largely studied in distributed IR [4]. In aggregated search, vertical selection has been treated as a query classification problem [2], [11], [8]. Some approaches focus on the integration of specific verticals such as products [11], news[8], or jobs[11]. Arguello et al. [3] show how this problem can be generalized to new verticals.

B. Visualizing: blended or unblended aggregated search

In aggregated search, there are two main approaches for results visualization, namely blended and unblended [14]. Blended visualization is the approach taken from major search engines such as Google where results from different verticals are merged and ranked in the same list. Unblended approaches place the results from each source in predefined panels. Here we can list as an example Yahoo! Alpha[1] or Kosmix [2]. Unblended approaches do not perform any source (vertical) selection as they always show results from all sources.

[1]http://au.alpha.yahoo.com/
[2]http://www.kosmix.com/

C. Evaluation

This section describes evaluation by regarding first AS interest and then the existing evaluation methodologies of AS approaches.

The interest of AS has been evaluated in [12], [2], [14]. It has been shown that vertical search intent is often present within Web search queries, which supports the interest in aggregated search. Liu et al. [12] analyzed 2153 generic Web queries into verticals, using query logs. They found that 12.3% have an image search intent, 8.5% have a video search intent and so on. Arguello et al. [2] classified 25195 unique queries, randomly sampled from search engine logs, into 18 verticals. 26% of the queries (mostly navigational) were assigned no vertical, 44% of the queries were assigned one vertical and the rest of the queries were assigned more than one vertical. The latter were mostly ambiguous.
Sushmita et al. [14] studied the quantity and diversity of collected relevant results by the use of an aggregated search interface, showing that this surpasses a traditional tabbed interface. Their evaluation involved 6 broad tasks and 16 participants.
Diversity has its advantages. Since early times it has been stated that a document should not only be relevant, but also novel [9], [5]. Diversification can maximize chances to guess at least one relevant result [1] and it can help users to collect multiple aspects of an information need [7].
The facts that vertical intent is often present within Web search queries and that aggregated search produces a diverse set of relevant results can prove partly the interest in aggregated search. However, if a vertical search engine does not produce relevant results for the query, aggregation is useless. Furthermore, we need to know how and why two different sources can contribute to answer the same information need. Do they provide duplicate information or not ?

Until now, different evaluation methodologies have been undertaken for evaluating the effectiveness of aggregated search. Some of them have been used to evaluate source selection approaches [2], [11], [12] and some others to compare aggregated search interfaces [14], [16], [17].
One common protocol is to ask human participants to choose which are the relevant sources for a query [2], [11], [12]. Queries in these works are not associated with any information need. As a result, the participant might not guess the real information need or might neglect some interpretations of the query and some queries might demand specific knowledge.
In [14], [16], Sushmita et al. compare the effectiveness of different interfaces for aggregated search. They show that users find more relevant results when vertical results are placed together with Web results. They also show that placing vertical results on top, on bottom or in the middle

of search results can affect the amount of vertical search results accessed by users. In both studies, participants are shown concrete search results from the considered sources. They are also given the information need behind the query. They have to click on results and bookmark the ones that are relevant. This approach is closer to traditional IR evaluation. The information need is not ambiguous and assessors can access real search results.

Instead of human participants, relevance assessments have been simulated using click-through logs [15], [8], [16]. In [8], Diaz shows that queries which obtain a high click through rate within news results are probable to be newsworthy. Click-through logs are also used in [16]. It is shown that for some sources such as video click-through behavior is different. Although click-through logs enable a large scale automatic evaluation, they cannot be as realistic as human based evaluation.

Whatever the aim of the presented works (evaluating the interest of AS or evaluating AS approaches), they all considered one type of relevance (intent or content) and one type of query (with or without a fixed need). Contrarily to these approaches, we will consider all 4 dimensions of study. The following section describes our experimental setup.

III. Experimental setup

We built a user study which aims at evaluating first the interest of AS and second the impact of relevance and query types on evaluation.

Our first objective is evaluated by considering the following points :

- analyze the distribution of relevant results across different sources,
- examine the interest of different vertical searches compared to traditional web retrieval,
- analyze why different sources are useful and if different sources are complementary to each other

Our second objective is to measure the impact on evaluation results that might be generated by relevance and query types.

The user study simulates 4 evaluation tasks which involve real participants evaluating relevance of 9 different sources. We consider two types of source relevance (by intent or content) and two types queries (with or without a fixed information).

A. Definition of source relevance

Intuitively, a relevant source is the one that can provide useful information for an information need. In this context, source relevance is different from search result relevance. A source might be relevant for a query (e.g. "sober Amy Winehouse photos" with respect to image search) even if no relevant results exist for this query. We can define source relevance with one the following statements:

- **Definition 1 - Relevance by intent:** a source is relevant for a query if it makes sense to issue this query to this source (i.e. there can be an intent for this source). The source does not even have to be concrete. For instance, the abstract source "image search" is relevant for the query "Eiffel tower photos". Relevance assessment does not depend on the quality of any search engine.
- **Definition 2 - Relevance by content:** a source is relevant for a query if it contains relevant results for the query. The source has to be an existing search engine and relevance assessment depends on the quality of the source and the availability of relevant results for the query.

B. Queries

We randomly sampled 100 queries from the Million Query Track in TREC 2007 [6], which contains itself queries extracted from search engine logs. This is done for the following reasons. First, choosing manually queries or writing down ourselves queries can bias results. Second, we could not find a collection of query logs free to use, which would be the ideal solution. The choice of the Million Query Track can be a bias as queries are likely to have textual intent[3]. However, this is not a major issue as our goal is to compare evaluation dimensions rather than the real distribution of query intents. Selected queries are available online at http://www.irit.fr/~Arlind.Kopliku/Evasion_queries.txt if experiments are to be reproduced.

In the following, we distinguish among **short-text query** and **query with fixed need**. A short-text query corresponds to a query as it is provided by TREC (1 to 3 words). We associated to each query a detailed description (a TREC-like textual description) corresponding to what we considered as relevant for this information need. This is what we call query with fixed need.

C. Sources and participants

We consider 9 sources: 8 vertical searches and Web search. The sources had to be diverse with minimal intersection. Most of them are common well known vertical search engines. They were chosen for two main reasons. First, they have already been used in state of the art approaches [2], [14], [15]. Second, we could find a free API for the source. The list of sources is the following: *Web search, Video search, Image search, News search, Maps (geographic search), Wikipedia search, Product search, Answer search, Definitions*.

For evaluating relevance by content, we used the following real sources:

- Yahoo! BOSS to get results for Web, image, news, answers and Wikipedia,
- API-s offered by Google Maps, Tuveo, E-Bay and Bing for geographic search, videos, products and definitions respectively.

[3]There is at least one match in the GOV2 collection.

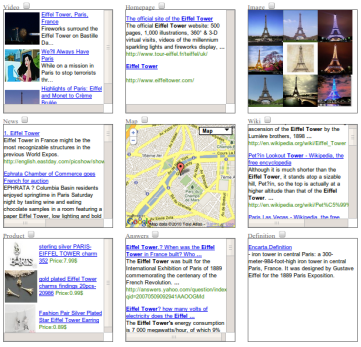

Figure 1. The interface for evaluating relevance by content with the results for the query "Eiffel tower"

The user study was conducted with 33 participants belonging to our research institute with 14 master students, 14 PhD students, 6 lecturers and 1 engineer. There were 13 women and 22 men. They all had a sufficient English level.

D. Evaluation interface

When evaluating relevance by content, we use a simple way to show search results which corresponds to an unblended search approach. This is done for the following reasons. First, we do not propose any aggregation algorithm that ranks results of different sources. Second, we want participants to access results from all sources without favoring any source. Results are shown in 9 panels (see figure 1) forming a 3x3 square. Each panel is labeled by its source and contains only results from that source (vertical search or Web search). To avoid biases, sources are shown in a random order for each query.
Participants might be skeptical about results coming from one source or they might expect results in one source and not in another accordingly to the issued query. To avoid this kind of bias results are shown in a homogeneous manner and the participants are instructed to view results from all sources. The user study is conceived to produce no bias among sources except of the results quality itself.
The number of results per source to be shown is chosen based on the visualization space and the comprehensibility of the results. For geographic search, only one result is shown in the map. The panel for images is filled with 9 images. For the rest of the sources 3 results are shown for each. It was possible to show 9 videos, but videos are not as self explanatory as images. Showing a title with some description is more informative.

Figure 1 presents an example of the interface shown to participants for the query "Eiffel tower".

E. Tasks description

We evaluated 4 search configurations named here tasks, each one corresponding to one type of relevance and one type of queries.

- **Task 1:** short-text query, relevance by intent
- **Task 2:** short-text query, relevance by content
- **Task 3:** fixed need, relevance by intent
- **Task 4:** fixed need, relevance by content

Those tasks are detailed below.

1) ***Task 1****:* This task is similar to evaluation in [2], [11], [12]. In this task, the participants have to decide for each short-text query which sources are likely to be relevant (relevance by intent). In other terms, they have to choose in which sources they expect to have useful results for that query. They are free to have all possible interpretations for the query. If they do not get any interpretation for the query they have to say so. Table I shows an example for the query *"London Tower"*. Different sources such as Web, Wikipedia and images can be relevant.

Query: *London Tower*				
Video	**Web**	**Image**	**News**	**Map**
No	Yes	Yes	No	Yes
Wiki	**Product**	**Answers**	**Definition**	
Yes	No	No	No	

Table I
EXAMPLE OF RELEVANT AND IRRELEVANT SOURCES FOR THE QUERY "LONDON TOWER"

2) ***Task 2****:* In this task, each short-text query is submitted to the 9 vertical searches. Participants are shown results from all sources (if some results exist for the considered source) in the interface presented in section III-D. This task is not meant to evaluate the way of presenting results (through the interface), but the utility of each source in the context of limited visualization space as well as the utility of the whole combination of sources and results.

A source is considered relevant if it returns at least one relevant result, whatever the interpretation of the query. The participant is indicated to view results from all sources, even if he has an interpretation in his mind and even if he does not expect relevant results for a given source.

3) ***Task 3****:* Here the participants are given a fixed need query (with the description of the information need that could have triggered the query). Using both the query and the information need, they have to choose the sources they guess to be relevant (relevance by intent).

4) ***Task 4****:* In this last task, the participants are also given a description of the information need that could have triggered the query as well as they are shown results from each source. This task is closer to traditional IR evaluation. As for Task 2, a source is considered relevant if it introduces at least one relevant result (relevance by content).
The participants are also asked to tell which is the most useful source for the information need. If they can choose one, they have to tell if they find this source enough for the information need. This allows us to investigate deeper the advantages of AS.

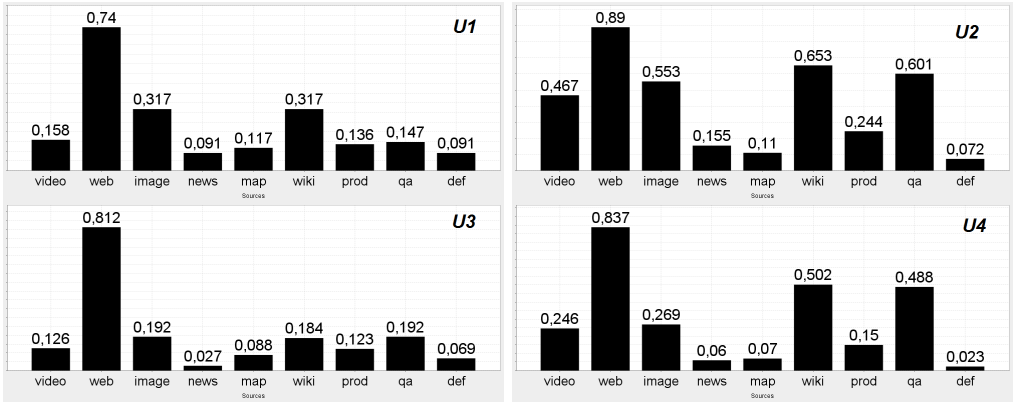

Figure 2. Relevance of sources based on $U1$ (top left), $U2$(top right), $U3$(bottom left) and $U4$(bottom right)

Study	Video	Web	Image	News	Map	Wiki	Prod	QA	Def	Total
U1	0.15	0.74	0.31	0.09	0.11	0.31	0.13	0.14	0.09	2.11
U2	0.46	0.89	0.55	0.15	0.11	0.65	0.24	0.60	0.07	3.74
U3	0.12	0.81	0.19	0.03	0.08	0.18	0.12	0.19	0.06	1.81
U4	0.24	0.83	0.26	0.06	0.07	0.50	0.15	0.48	0.02	2.64

Table II
COMPARATIVE TABLE OF THE AVERAGE RELEVANCE OF SOURCES WITH RESPECT TO TASKS

F. Task deployment

All tasks except the second involve 10 participants with 30 queries each. The second task involved 30 participants with 10 queries each. This was done as task 2 takes longer than the other studies. A participant who participated in all tasks would evaluate 100 queries. Tasks were designed to have every query assessed by exactly 3 participants, therefore we have 300 queries assessed by task.

IV. RESULTS

In this section we first discuss the results concerning the interest of aggregated search and then we compare evaluation results by considering different types of relevance and queries.

Before giving these results, we list in the following paragraph the notations we used.

A. Notation

In the figures, graphs and tables of this section, we will use the following abbreviation for the sources: *video for video search, web for web search, wiki for Wikipedia, map for geographic search, prod for product search, image for image search, def for definitions, qa for answer search*. We will use $U1$, $U2$, $U3$ and $U4$ to refer respectively to task 1, task 2, task 3 and task 4 of our user study.
Furthermore, we need some definitions. Let q_i^u be the i^{th} query assessed by participant u. Let Q be the set of all q_i^u in a task, $Q = 300$ for each task.

Finally, let $Q(s_j)$ be the set of q_i^u where the source s_j is considered relevant.

B. Interest of aggregated search

Let us define the *average relevance of a source* s as the proportion of queries for which s was relevant. Relevance might be assessed by intent or content depending on the task being considered. We will denote it as $R(s)$ and it corresponds to:

$$R(s) = \frac{|Q(s)|}{|Q|} \tag{1}$$

Results here will be discussed according to the goals mentioned at the beginning of section III.

Table II compares the average relevance of sources with respect to the different tasks. The same data is also shown in figure 2 to ease comparison between different sources. We can see that the Web source is the most relevant and definitions source is the less relevant whatever the considered relevance or query type. If we order sources by average relevance, not all tasks produce the same ranking. However all 4 tasks confirm that **relevance is sparse across all sources**.

We also compare the relevance of vertical searches with Web search with respect to the number of queries that can be answered respectively by Web search and the union of all verticals. Let A the set of all sources, *web* be the Web search, V be the set of all verticals *(V=A-web)*. We will say that $R(A)$ represents the proportion of queries where there is at least one relevant source, $R(V)$ represents the proportion of queries where there is at least one relevant vertical. We also denote as $R(web, \overline{V})$ the proportion of queries where the Web search is relevant alone (no relevant vertical) and $R(V, \overline{web})$ the proportion of queries where there is at least

one relevant vertical but Web search is considered irrelevant. Results are shown in table III. The last column corresponds to an average over all tasks.

Study	U1	U2	U3	U4	avg
$R(A)$	0.99	1.00	0.99	0.97	0.98
$R(web)$	0.74	0.89	0.81	0.83	0.82
$R(V)$	0.75	0.92	0.66	0.79	0.78
$R(web, \overline{V})$	0.24	0.08	0.31	0.17	0.19
$R(V, \overline{web})$	0.25	0.11	0.18	0.13	0.16

Table III
DURATION AND AGREEMENT BY USER STUDY

We notice (in row $R(A)$ of the table) that about 98% of queries have at least one relevant source. Web search is relevant for about 82% of queries, while 78% of queries have at least one relevant vertical. Furthermore, Web search is relevant alone (no relevant vertical) for only about 19% of queries. The interesting result comes from the last raw of the table where, where for about 16% of the queries have at least one relevant vertical without having Web search as relevant. We can draw the following conclusions. **Vertical sources can answer many queries. Mostly they are relevant at the same time with Web search, but they can also present relevant results when Web search fails to.**

Let us now examine if many sources can be relevant at the same time and if sources are complementary with each other. Concretely, we want to analyze how often multiple sources are considered relevant and when this is the case we want to see if multiple relevant sources provide same information (same results) or they return different results for a given query.

Before discussing deeply this question, we first counted the number of queries that have more than one relevant source. We found 155 queries for task 1, 251, 134 and 215 for tasks 2, 3 and 4 respectively. There are also some queries with 4, 5, 6 even 7 relevant sources. The average number of relevant sources per query and per task is in fact listed in the last column of table II. It is respectively 2.11, 3.74, 1.81 and 2.64 for the tasks $U1$, $U2$, $U3$ and $U4$. We can therefore conclude that **many queries match more than one relevant source**.

We then attempted to understand the reasons of having more than one relevant source for a given query. We identified three major reasons. First, the query can be ambiguous (e.g. jaguar (car vs. animal)). Second, the query can be broad and it can represent an information need that demands multiple aspects (e.g. visit France can demand for maps, images, Web, etc.). Third, two different sources can return the same information. Although we chose a diverse set of sources, there are cases when two different sources present the same information. The most frequent is Web search which can intersect with all other sources.

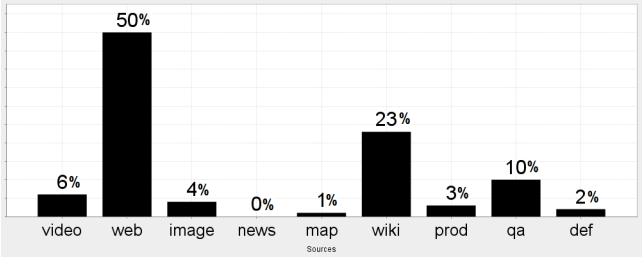

Figure 3. Distribution of the primary sources

We investigated further the last two explications. We examined whether the sources return mostly same results (same information) for a query or contrarily, they tend to return additional information that help to better answer the query, we call here complementary results. For this purpose, participants were asked in task 4 to select the most relevant source (called primary source) and to tell if this source was sufficient to satisfy the information need or if the other sources help, if they complementary?".

Among 300 queries (10 participants x 30 queries), in 272 cases a most useful source was selected. Figure 3 shows the distribution of primary sources. One notices that the Web is the most useful source half of the times, for the other half it is the other sources that present the most useful information. For this half, the vertical search is more directly relevant to the information need than the Web search.

The most useful source alone was considered sufficient for the information need 106 times (39%) and it is insufficient 166 times (61%). If the source is not enough it is necessary to provide additional complementary information. Apparently, this is needed quite often. The participant was also asked to tell if he/she finds additional useful information in other sources for his information need. For the 106 cases where the most useful source was considered sufficient alone, an average of 2.13 other sources were selected as useful. We can say that the participant found the primary enough for his information need, but at his/her surprise he could meet more relevant information in the other sources. For the 166 cases where the most useful source was considered insufficient alone, an average of 2.74 other sources were selected as useful. We can expect these sources to provide complementary information which is missing in the primary source. **This shows implicitly that aggregated search is useful because often one source alone is not enough, while multiple sources can complement each other for a completer answer.**

C. Impact of relevance and query types on evaluation

The four tasks of our user study can be used to evaluate interest in AS, but also to evaluate different approaches of AS. Until now, there is no common agreement evaluation technique for aggregated search. We will analyze aggregated search evaluation using the 4 techniques presented in this

paper.

1) Short text query vs. fixed information need: Using table II we see that fixing the information need decreases the amount and diversity of relevant sources. This can be seen comparing results from task 1 and task 3, but also by comparing results from user task 2 and user task 4. For instance, for the query *"Hamilton County"*, the information need was *"I want the location of Hamilton County"*. For this query, in $U1$ we can choose images and Wikipedia as relevant. Once fixed the information need these sources become irrelevant.

2) Relevance by intent versus relevance by content: Our aim here is to discuss the differences of results when considering relevance by intent or by content.

We found that both types of relevance assessment contribute in identifying relevant sources. However, there are sources that are considered relevant by intent and irrelevant by content and vice-versa. Concretely, in 7.2% of cases a source was considered relevant by intent (by at least one participant) in U1 and irrelevant (by content) in U2 (from all participants). In 26.3% of cases, a source was found relevant in $U2$ and irrelevant in $U1$. In 7.3% of cases, a source was found relevant in $U3$ and irrelevant in $U4$. In 19.6% of cases, a source was found relevant in $U4$ and irrelevant in $U3$. **Clearly, participants could identify more relevant sources by content than by intent**.

The above has some explications. For many queries, participants did not have enough knowledge or they could not imagine any relevant information in some specific source. Viewing results helps them finding additional useful information.

On the other hand, some sources can be considered irrelevant by content because they provide no relevant results, even if the participant could imagine something relevant by intent ($U1$ and $U3$). There can be at least two explications. There exists nothing relevant for that query or the source did not work well.

In order to fully understand these results, we analyzed the participants agreement as well as time they spent assessing the results. This analysis is described below.

3) Participants agreement and time: Let session duration be the average time needed for one participant to assess a query for all sources. The session durations are shown in table IV. As we can expect, evaluation of relevance by intent is 5 (fixed need) to 8 (query alone) times faster than evaluation by content.

We also computed participant agreement for each user study. We use Fleiss' Kappa k, a measure for assessing the reliability between a fixed number of participants. It is used in alternative to Cohen's Kappa which is used to measure agreement between exactly two participants. Depending on the range this measure falls, we can classify agreement into 6 classes: poor agreement ($k < 0$), slight agreement ($k \in [0, 0.2]$), fair agreement ($k \in [0.21, 0.4]$), moderate agreement ($k \in [0.41, 0.6]$), substantial agreement ($k \in [0.61, 0.8]$) and almost perfect agreement ($k \in [0.81, 1]$). Results are shown in table IV. The first user study falls in the range of fair agreement. The others fall in the range of moderate agreement. We can see that fixing the information need increases agreement and that greement is highest for evaluation by content.

The agreement level of our tasks does not affect the validity of results in the previous sections, because low inter-assessor agreement is common in IR evaluation. This has also been recognized in the context of major evaluation campaigns such as TREC, INEX [10].

Study	U1	U2	U3	U4
Session duration (seconds)	25	180	24	105
Inter participant agreement	0.36	0.48	0.46	0.56

Table IV
DURATION AND AGREEMENT BY USER STUDY

To conclude, we showed that evaluating by intent can introduce bias in the evaluation process. Participants miss many relevant sources, which thay can identify when they access results by content. Evaluating by intent is faster (see table IV), but it produces a lower participant agreement.

V. DISCUSSION

Our study proves that aggregated search is not just about coloring search results with images and videos. We could identify clear advantages of aggregated search. First we show that relevant results are present or expected in a diverse set of sources. Vertical search engines can answer many information needs. Mostly they are relevant at the same time with Web search, but they can also present relevant results when Web search fails to.

Vertical search engines return information that can be found in the Web, but they have a special focus/domain. They are supposed to work better when the query falls in their domain. Not only, we notice that vertical intent is frequently present in queries, but we also show that vertical search can answer these queries even where Web search fails.

A further investigation concerns diversity of relevant results. We find that many queries match more than one relevant source. We could identify three reasons to explain the latter. First, some of our queries are ambiguous. Second, some sources complete each other (return different aspects of the same need). Third, some sources return repeating information. We show that often one source alone is not enough, while multiple sources can complement each other for a completer answer. At our knowledge, no investigation existed until now to explain the diversity of relevant sources.

We investigated the differences among the 4 search situations trying to derive useful conclusions for the evaluation of AS. First, we notice that fixing the information need decreases the amount and diversity of relevant sources, but it eases the evaluation task. Less assessment time is needed and inter-assessor agreement is higher.
Second, we notice that assessing by intent has some significant drawbacks. Assessors miss interpretations of the query. They can identify more relevant sources when they access search results (relevance by content). The major strength of this kind of evaluation is assessment time which is significantly lower.
Task 1 of our study uses the same configuration as evaluation in [2], [11], [12] (short queries, relevance by intent). Although this form of evaluation is fast, it is also the less realistic. Assessors do not know the real information need and they often miss identifying relevant sources. On the other hand, evaluation as done in [14], [16] is closer to traditional IR evaluation (fixed need, relevance by content). The major drawbacks of this approach is that it is time-consuming and that it depends on the quality of sources.

The learnings in this paper might be useful for future evaluation of AS approaches.

VI. Conclusions

In this paper we investigated the interest and the evaluation techniques for aggregated search. We considered both short-text queries and fixed need queries. We also investigated for the first time both relevance by intent and relevance by content. In each task of the user study, relevance is assessed for 100 queries across 9 sources (vertical searches and Web search).
We could identify advantages of aggregated search which were not assessed before. Furthermore comparing the 4 search situations we could derive useful thoughts on the evaluation of aggregated search.

References

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14, New York, NY, USA, 2009. ACM.

[2] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference*, pages 315–322, New York, NY, USA, 2009. ACM.

[3] J. Arguello, F. Diaz, and J.-F. Paiement. Vertical selection in the presence of unlabeled verticals. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference*, pages 691–698, New York, NY, USA, 2010. ACM.

[4] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference(Special Issue of the SIGIR Forum)*, pages 21–28. ACM Press, 1995.

[5] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA, 1998. ACM.

[6] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. Evaluation over thousands of queries. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 651–658, New York, NY, USA, 2008. ACM.

[7] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, pages 659–666, 2008.

[8] F. Diaz. Integration of news content into web results. In *WSDM*, pages 182–191, 2009.

[9] W. Goffman. A searching procedure for information retrieval. *Information Storage and Retrieval*, 2(2):73–78, 1964.

[10] B. Larsen, S. Malik, and A. Tombros. Focused access to xml documents. chapter A Comparison of Interactive and Ad-Hoc Relevance Assessments, pages 348–358. Springer-Verlag, Berlin, Heidelberg, 2008.

[11] X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346, New York, NY, USA, 2008. ACM.

[12] N. Liu, J. Yan, and Z. Chen. A probabilistic model based approach for blended search. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 1075–1076, New York, NY, USA, 2009. ACM.

[13] V. Murdock and M. Lalmas. Workshop on aggregated search. *SIGIR Forum*, 42(2):80–83, 2008.

[14] S. Sushmita, H. Joho, and M. Lalmas. A task-based evaluation of an aggregated search interface. In *SPIRE '09: Proceedings of the 16th International Symposium on String Processing and Information Retrieval*, pages 322–333, Berlin, Heidelberg, 2009. Springer-Verlag.

[15] S. Sushmita, H. Joho, M. Lalmas, and J. M. Jose. Understanding domain relevance in web search. In *WWW 2009 Workshop on Web Search Result Summarization and Presentation, Madrid*, 2009.

[16] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. Factors affecting click-through behavior in aggregated search interfaces. In *CIKM '10: Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 519–528, New York, NY, USA, 2010. ACM.

[17] P. Thomas, K. Noack, and C. Paris. Evaluating interfaces for government metasearch. In *IIiX*, pages 65–74, New York, NY, USA, 2010. ACM.