Aggregated search: From information nuggets to aggregated documents

Arlind Kopliku

Institute de Recherche en Informatique de Toulouse, UMR 5505 CNRS, SIG-RFI
118 route de Narbonne F-31062 Toulouse Cedex 9
Arlind.Kopliku@irit.fr

ABSTRACT. *The aggregated search assembles in one interface information from different sources. It deals with different types of content (text, video, image, etc) and granularities of retrieval. It aims to assemble the retrieved content forming an aggregated result. This is in contrast with the common approach which provides a list of documents as the search result.*

Today we are able to retrieve content of different types and granularities, but little work has been done for their aggregation. Being a new area of research formalization of aggregated search is still missing. This paper treats the problem in a high level of abstraction. It presents the state of the art and decomposes the problem, listing issues and providing examples and formalization. This work aims to be a base of reflection and a reference for future work.

RÉSUMÉ. *Le but de la recherche agregée est de rassembler des informations provenant de plusieurs sources en une seule interface. Elle doit ainsi gérer des problématiques liées aux différents types de contenu (texte, vidéo, image, etc) ainsi qu'à la granularité des résultats. La formation d'un contenu agrégé à partir de différents types de contenus retrouvés contraste avec l'approche commune en RI consistant à renvoyer à l'utilisateur une liste ordonnée de résultats. Si nous sommes aujourd'hui capables de retrouver de l'information de différents types et de différente granularité, très peu de travaux existent concernant leur agrégation. La recherche agrégée étant un domaine de recherche récent, elle manque encore de formalisation. Ce papier se propose de traiter la recherche agrégée à un niveau d'abstraction élevé. Il présente tout d'abord l'état de l'art, puis décompose le problème en listant et formalisant les différentes problématiques. Ce travail doit servir de base de réflexion et de référence pour de futurs travaux sur le domaine.*

KEYWORDS: *aggregation, aggregated search, unified search*

MOTS-CLÉS : *agrégation, recherche agrégée, recherche unifée*

1. Introduction

Aggregated search comes in contrast with classical search paradigm, where search systems list a number of documents as an answer to a free language query. In the latter, the user has to scan the returned list and search within one or some of the retrieved documents. This can be time consuming especially if the needed information is only a small portion of a document or when the needed information resides in more documents. Aggregated search proposes fictitious built documents which should contain, organize and connect useful information.

Today we are able to retrieve information of different types and at different granularity such as paragraph, chapter, document, etc, but there is little work providing means to combine them. The information need can be composed of sections of text belonging to different sources as well as it can contain some images, videos, etc. Aggregated search tries to identify the necessary content, organize it and present it to the user in such a way to facilitate his information search. Aggregation can also provide a richer view of the existing information.

Existing solutions are very limited. However, some search engines have already started to merge information of different types into the main result search page. The contents are simply listed or placed into fixed visualization spaces. To our knowledge, scientific publications are even more limited. Most of the work is at the proposal level or it is about specific issues in specific contexts.

Our aim in this paper is not to provide a solution, neither to list existing ones. Aggregated search is very recent and needs formalization. Thus, we aim to present the problem in general and decompose it, listing the main issues. This way we provide a mean of reflection and reference for future work.

This paper is organized as follows. Section 2 describes the state of the art and why a lot of work is to be done. Section 3 presents the aggregation search problem. It describes the aggregation process providing the necessary phases and it provides examples and reflections on the types of aggregation. Section 4 is about conclusions and future work.

2. State of the art

Aggregated search intersects many areas. It starts from the retrieval of contents and it ends with their filtering, selection and organization. A lot of work exists on the retrieval side. But content aggregation has not been largely studied. A lot of work is to be done and formalization is almost missing.

A good study starting point can be focused retrieval (Fuhr *et al.*, 2008). Focused retrieval deals with the granularity issue. Aggregated search could use it to provide the input to assemble and organize. In fact, in the XML retrieval context it has been shown than returning several elements together trigger a stronger user satisfaction than returning a single element. An interesting case study comes from the INEX Relevant in Context task(Fuhr *et al.*, 2008). Here, instead of returning elements separately, relevant elements are grouped by document. Still, it does not consider grouping from different sources.

During the ACM SIGIR 08 conference (Lalmas *et al.*, 2008), a workshop was held especially for aggregated search. Sushmita, Lalmas, Tombros (S.Sushmita *et al.*, 2008) propose to visualize as search results, digest pages which are thought as fictitious documents built from clustering the documents returned by a search engine. Some others focus in specic domains such as social science and medicine (Ou *et al.*, 2008; Wan *et al.*, 2008). The first one considers a collection of social science articles. It extracts and organizes important concepts. The second article focuses on the importance of result organization to the user utility.

The industry already comprises aggregated search features. We can find it in specific contexts such as product search or location search. For example, *Wize.com* offers product search with results that are obtained as an aggregation from several sources (Shilman, 2008). Google's local search[1] adds phone numbers, images and web pages when available in addition to the map result.

Aggregation seems to be the trend in web search, too. This trend is often referred as *unified search* , but other names are in use such as *blended vertical search* or *universal search*. The new approach adds to the list of web pages presented in the main search results page vertical content such as maps, images, news, etc. There are two main approaches at this moment. In the first one, the content appears inline with the HTML listings and all of it is ranked according to a scoring algorithm. Google Universal search reflcts this approach. The other approach involves a new search results layout with standardized "holes" for each type of content. *Ask3D* [2] , *Kosmix* [3], *Yahoo's Alpha*[4] and *Google's SearchMash*[5] are all taking this approach.

Existing approaches remain very limited. In general, almost all solutions predefine the way the content should be organized: some use predefined content placement and some other simply order by relevance. Relations between the retrieved contents are not considered. However, sometimes some information has to be shown before another even if it is less relevant (logical connection, chronological order, etc). Moreover, some contents should be grouped together (similar content, alternative lecture, A explains B, A is a photo of B, etc). Studying relationship between results would not only help better visualize the contents, but also provide supplementary useful information which is not deemed as relevant in the beginning.

[1] http://maps.google.com
[2] http://www.ask.com
[3] http://www.kosmix.com
[4] http://au.alpha.yahoo.com/
[5] http://www.searchmash.com

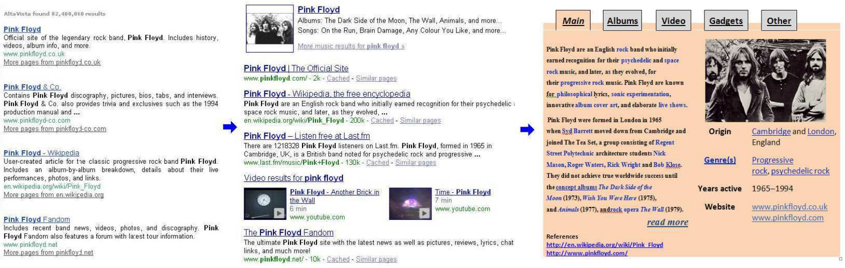
3. From information nuggets to aggregated documents

3.1. *Definition*

In the ACM SIGIR 2008 workshop on aggregated search(Lalmas *et al.*, 2008) the following definition was given:

Definition *Aggregated search is the task of searching and assembling information from a variety of sources and placing it in a single interface.*

In fact, the user might be satisfied with a part of a document or some parts. Sometimes, his information need might be the composition of some sections from different documents. He might also want some images, videos, etc. Let's consider an example with the query "Pink Floyd". The classical ad-hoc approach would list a list of web pages, probably repetitive and partially relevant. Google's universal search would list contents of different types augmenting the diversity of information. Aggregated search would try to deal not only with redundancy and diversity but it would also organize the information. It can start with a description of the group, general data, some images, videos, albums, etc. Figure 1 better illustrates the approaches.



AltaVista found 82,400,000 results

Pink Floyd
Official site of the legendary rock band, **Pink Floyd**. Includes history, videos, album info, and more.
www.pinkfloyd.co.uk
More pages from pinkfloyd.co.uk

Pink Floyd & Co.
Contains **Pink Floyd** discography, pictures, bios, tabs, and interviews. **Pink Floyd** & Co. also provides trivia and exclusives such as the 1994 production manual and **...**
www.pinkfloyd-co.com
More pages from pinkfloyd-co.com

Pink Floyd - Wikipedia
User-created article for the classic progressive rock band **Pink Floyd**. Includes an album-by-album breakdown, details about their live performances, photos, and links.
en.wikipedia.org/wiki/Pink_Floyd
More pages from en.wikipedia.org

Pink Floyd Fandom
Includes recent band news, videos, photos, and discography. **Pink Floyd** Fandom also features a forum with latest tour information.
www.pinkfloyd.net
More pages from pinkfloyd.net

Pink Floyd
Albums: The Dark Side of the Moon, The Wall, Animals, and more...
Songs: On the Run, Brain Damage, Any Colour You Like, and more...
More music results for **pink floyd** »

Pink Floyd | The Official Site
www.**pinkfloyd**.com/ - 2k - Cached - Similar pages

Pink Floyd - Wikipedia, the free encyclopedia
Pink Floyd are an English rock band who initially earned recognition for their psychedelic space rock music, and later, as they evolved, **...**
en.wikipedia.org/wiki/**Pink_Floyd** - 200k - Cached - Similar pages

Pink Floyd – Listen free at Last.fm
There are 1218328 **Pink Floyd** listeners on Last.fm. **Pink Floyd**, formed in 1965 in Cambridge, UK, is a British band noted for psychedelic rock and progressive **...**
www.last.fm/music/**Pink+Floyd** - 130k - Cached - Similar pages

Video results for **pink floyd**

Pink Floyd - Another Brick in the Wall
6 min
www.youtube.com

Time - Pink Floyd
7 min
www.youtube.com

The Pink Floyd Fandom
The ultimate **Pink Floyd** site with the latest news as well as pictures, reviews, lyrics, chat links, and much more!
www.**pinkfloyd**.net/ - 10k - Cached - Similar pages

Pink Floyd are an English rock band who initially earned recognition for their psychedelic and space rock music, and later, as they evolved, for their progressive rock music. Pink Floyd are known for philosophical lyrics, sonic experimentation, innovative album cover art, and elaborate live shows.

Pink Floyd were formed in London in 1965 when Syd Barrett moved down from Cambridge and joined The Tea Set, a group consisting of Regent Street Polytechnic architecture students Nick Mason, Roger Waters, Rick Wright and Bob Klose. They did not achieve true worldwide success until the concept albums *The Dark Side of the Moon* (1973), *Wish You Were Here* (1975), and *Animals* (1977), androck opera *The Wall* (1979).

read more

Origin	Cambridge and London, England
Genre(s)	Progressive rock, psychedelic rock
Years active	1965–1994
Website	www.pinkfloyd.co.uk www.pinkfloyd.com

References
http://en.wikipedia.org/wiki/Pink_Floyd
http://www.pinkfloyd.com/

a) Ad-hoc retrieaval(Altavista)

b) Unified search (Google's approach)

c) Aggregated search

Figure 1. *Comparing search paradigms assuming the query "Pink Floyd" for all three cases. The aggregated result is hand built.*

Aggregated search does not look for single documents. It tries to build them. It merges content together. We will define the aggregated document as its search objective.

Definition *An aggregated document is the result of assembling information from a variety of sources.*

If aggregated documents are built by aggregation, we are interested in their components. We will categorize these components by type. The term nugget will represent a component of a certain type. The following definition better describes it.

Definition *The information nugget is a unit of information of one type of content (text, image, video, news, etc).*

In the context of textual information, it can represent a partition of the text. In the case of images, it is the image itself. In the case of video, it can be the entire video or a part of it. It is clear now that we want to build aggregated documents assembling information nuggets.

3.2. *Aggregate structure*

Aggregated search has to deal with the organization of relevant content. It is not enough to gather the information but it is also fundamental to organize it for its final visualization to the user. We would like to define as aggregate structure all the information which describes the content, the order of visualization,and the preferences in visualization of the aggregated document. We want to keep out the aesthetic considerations such as background color or font size.

Definition *The aggregate structure describes the content and visualization preferences of the aggregated document. Stylistic considerations (such as background color, font size) are left out.*

The following examples illustrate some partial structure of some aggregated information:

- –3 images, 2 videos
- –a list of 4 reviews, 3 ratings
- –passage A, passage B, passage C, visualization order A,B,C

In ad-hoc information retrieval the answer to a query is a list of documents. In aggregated search it should be possible to add aggregate structural information to the query. Someone could ask for images and videos only, or reviews and ratings only. But user studies (Nielsen, 2003) tell users are generally lazy. Though, they do not use additional options. Nevertheless, this can be very useful in times. Imagine a travel service in the web who wants to search for hotels and it is interested in hotels with at least 3 photos, 1 map, the address, phone number, number of stars, reviews, etc.

3.3. *Abstract phases*

In this section we define 3 abstract components of the aggregated search. They do not have to be implemented separately. They can be merged and ordered in many ways. Nevertheless, we consider they are part of the process. We present a selection, a filtering and an organization phase.

The **selection phase** is about selecting the information which is potentially useful. There are many possibilities. The list of the K top documents returned from a search engine is one choice.

The **filtering phase** is about filtering useless information. The selection and filtering phase correspond to the intuition that the aggregated search might show the most useful information and the less useless information. Filtering can be applied before selection. But, it seems more plausible placing it after selection or merging them together. Filtering could remove information redundancy. It could also deal with ads.

The aggregated search allows assembling information coming from different sources. Because the aggregated result components can be parts of documents, we loose the original structure. Here comes the **organization phase**. It is necessary to relate the retrieved content. The retrieved content has to be organized in order to be visualized in a sensed manner. Some content is more sensed to be listed; some has to be grouped and so on. Organization information can be also gathered at selection or filtering time.

3.4. *Types of content*

It is important to distinguish between different types of content because the retrieval and the aggregation process depend strongly on the types of content. There are many choices. Some types include some others and some times there is intersection. Below there is a brief list of types of content:

By type of media: text, video, image

Units of information: book, article, chapter, title, web page, word

Contextual use: blog, review, rating

Field: field, email, phone number, address

Link: anchor, hyperlink, reference

We can distinguish between media types such as text, video, image and hypertext. But we can also drill into different granularities within the same media such as book, chapter and paragraph in text. The type of information can be used in some cases such as for news, reviews, comments, etc.
We also define a special type of content field. Fields are composed of the filed name and the field value. Emails, phone numbers are examples of fields.

3.5. *Aggregation*

The aggregated result can be considered as composed of information (set of nuggets) and organization (aggregate structure). It is necessary though to be able to relate the retrieved content for the visualization. We will define first a list of logical relations, which can later be combined for visualization. Below, there is a short list of relation which is not exhaustive.
Association: We have an association when a content is related with another. Because the final aim is visualization, this property should reflect the probability of two or more contents to appear together. Association should try to define almost certain

relations between two entities like for example: this map is about this hotel.
Group: Some items share some properties. These properties can be used to group the content. We can put in the same group information of the same context. We can also group based to the type of content. Grouping can be done at different levels. It can be used for visualization, but as for association it does not force a specific visualization.
Order: Some content should be ordered. The ordering criteria can be various. We can order some content in chronological order, by relevance, etc. Relations of order give a preference to one item with respect to other items.
We can now define the most common types of visualization and how the logical relations can be used here. Visualization relations are part of the aggregation structure. They define visualization preferences. They can also be merged with each other.
List: Lists are a frequent type of visualization. The content here is simply listed in a consecutive order. The items might be ordered by some property or be presented randomly. Depending on the case, we can use the order or group logical relations. The listing can be vertical or horizontal.
Table: The tables are often used for visualization. They can be used for some logical organization of the data as well as for aesthetic considerations.
Block: We define as a block the visualization of content within a rectangular surface. The structure within can be a table. Blocks can be built using the group logical relation as well as some other properties. We can put within the same block information of the same context or same type of content, etc. Blocks can also be nested together or listed.
Links: Links can contain a brief description of a content and a link. They are very useful in information retrieval.
Menu: Menus are linked content by a list of links. Menus are useful to organize the visualization space. They can be viewed as linked blocks.
Partial inclusion: Sometimes an information can be very long. We might want to partially introduce it. This way the user can decide if reading more or not. A *"read more"* option can be provided or a link to the source.
Summary: A summary is a shortened version of the original source. The main purpose of such a simplification is to highlight the major points from the genuine (much longer) subject, e.g. one or more documents, a movie, an event, etc.

4. Conclusions and future work

This paper addresses aggregated search as a prominent area of interest. It presents the problem in general and then decomposes it trying not to lose generality. We provide phases and considerations necessary to build such a system. Although things are moving at enterprise level, there is still few published scientific research addressing aggregated search directly. That is why this paper contributes by presenting the problem definition, the actual state of art and identifying some of the issues and subproblems.

The paper starts at a high level of abstraction then it drills down into lower level. It proposes three abstract phases as essential components of aggregated search namely selection, filtering and organization. But, no constraints and specific implementations are proposed. There are in fact many possible solutions. Proposing a part of them is not the goal of this article which aims generality.

Aggregated search arises from the availability of information of different types and sources. Different types of content are retrieved and assembled differently. Here, we present an overview of the possible aggregations and we provide some types of relations which can be used to organize the aggregated result. This should help to reflect on specific solutions.

To solve the aggregated search problem, the following research issues can be considered. Focused retrieval and vertical searches can be used to feed the input to the system. Further filtering can involve redundancy removal, such as near-duplicate detection. Query interpretation is essential to understand the user need, but it can also indicate some aggregation hints. In order to aggregate the retrieved content, it is also necessary the study of the relations between content as well as the study of intelligent content organization (placement) algorithms. Finally, evaluation is also an open issue. User studies can be used but they are time consuming and not preferable. Devising a good evaluation system and realizing evaluation benchmarks are important challenges for the domain.

5. References

Fuhr N., Kamps J., Lalmas M., Malik S., Trotman A., "Overview of the INEX 2007 Ad Hoc Track", pp. 1–23, 2008.

Lalmas M., Murdock V. (eds), *SIGIR Workshop on Aggregated Search*, ACM, 2008.

Manning C. D., Raghavan P., Schutze H., *Introduction to Information Retrieval*, Cambridge University Press, 2008.

Nielsen J., *Designing Web Usability*, 9th edn, Dwyer, David, Indianapolis (USA), 2003.

Ou S., Khoo C. S. G., "Aggregating search results for social science by extracting and organizing research concepts and relations", *in* Lalmas *et al.* (2008), 2008.

Shilman M., "Aggregate documents: making sense of a patchwork of topical documents", *DocEng '08: Proceeding of the eighth ACM symposium on Document engineering*, ACM, New York, NY, USA, pp. 3–7, 2008.

S.Sushmita, M.Lalmas, A.Tombros, "Using digest pages to increase user result space: Preliminary design", *in* Lalmas *et al.* (2008), 2008.

Wan S., Paris C., Krumpholz A., "From Aggravated to Aggregated Search: Improving Utility through Coherent Organisations of an Answer Space", *in* Lalmas *et al.* (2008), 2008.