Aggregated Search Techniques Usability

Kamal Bal[1], Youssef Amghar[2], Adbessamed Réda Ghoamri[3] and Hakima Mellah[1]

[1] CERIST, Research Center on scientific and technical information, Algiers, Algeria

[2] INSA (Institut National des Sciences Appliquées), Lyon, France

[3] ESI (National Superior School of computer science), Algiers, Algeria

Abstract. The complexity issued from information seeking requires providing end-users with tools in order to obtain relevant information from heterogeneous sources and organize these pieces of information in an understandable and a coherent way. Aggregated search is a new information retrieval paradigm that aims to gather information from different sources and present them in a single interface. In this paper, we study the different aggregation techniques used in this information retrieval context and attempt to determine suitable use situation for each aggregation technique.

1 Introduction

Current information environments are characterized by the multiplication and the diversity of information sources [24]. Users need to extract relevant information from a large amount of information. Constant efforts are made by researchers and search engine companies in order to make information effectively accessible [24]. They aim to provide users with tools to obtain relevant information from several heterogeneous sources and organize retrieved information in an understandable and coherent way.

Information Retrieval Systems return a list of potentially relevant documents. Relevant information can be found entirely in a document or be scattered in several documents, which can also be derived from several sources [19]. The user has to ask the various sources, read through returned documents, select and organize relevant pieces of information to build an appropriate response [14]. These pieces of information can be of different granularities and different modalities (an image, a text, a video, or even an attribute with its value) [18].

A better solution is to construct the response by automatically selecting and organizing the different relevant information granules. This approach represents a new paradigm in multi-sources information retrieval called "Aggregated search". Aggregated search has been defined for the first time in SIGIR'2008 workshop [1]: *"it is a task trying to gather information from different sources and present them in a single interface*". In other words, aggregated search attempt to identify the relevant content, organize and present it to the user. Information of different types (text, image, video, etc...), and of different granularities (text passage, entities, attributes, etc ...) are connected and combined to compose an aggregated result [3].

Bal K., Amghar Y., Réda Ghoamri A. and Mellah H..
Aggregated Search Techniques Usability.
DOI: 10.5220/0004600000650073
In *Proceedings of the 2nd International Workshop on Web Intelligence* (WEBI-2013), pages 65-73
ISBN: 978-989-8565-63-1

Aggregated search has not been tackled as a whole, but many developed solutions involve aggregation features and many works can be seen as specific cases [3, 18]. In this paper we present and study the different aggregation techniques used in this information retrieval paradigm. The study of the techniques is done in terms of use in order to characterize suitable use situations for each technique.

The remainder of this paper is organized as follow: In section II, we present aggregated search paradigm and its specificities. Than we describe in section III, the different aggregation techniques used in aggregated search. In Section IV, we study theses techniques and try to identify some criteria characterizing the use of each technique. We conclude in section V.

2 Aggregated Search Paradigm

Aggregated search has been defined for the first time in SIGIR'2008 workshop in [1], as a task trying to gather information from different sources and presents them in a single interface. Aggregated search aims to search information in multiple information sources. Even if the paradigm of aggregated search is recent, the multi-source information retrieval exists for long time. Federated information retrieval [4] and meta-search engines [9] are search paradigms designed to provide search results from multiple sources. However, the information sources in aggregated search are heterogeneous and contain documents of different types of (text, images, news, video…) and different granularities. In addition, in aggregated search, information sources are managed by dedicated search engines, all under the main search engine, not several independent search engines, as is the case with meta-search engines. Jaime Arguello refers to these properties as *results type* and *retrieval algorithm heterogeneity* [23]. These properties have been ignored in anterior multi-sources information retrieval paradigms.

Content aggregation takes a great importance in aggregated search. Indeed, it is not enough to merge results lists of the same type to produce a single list, but it is to compare heterogeneous results, to decide how to aggregate and present them in a single interface. In 2000, the Korean retrieval engine Naver[1] introduced "*comprehensive search*" and began to incorporate multimedia answers in default search results. Google explicitly introduces the idea of aggregated search in 2007 via the concept of "*Universal Search*". Google searches through all its sources, compare and rank all the information in real time, and provide a unique and integrated search results page.

3 Aggregated Search Techniques

Aggregated search has not been treated as a whole, but many developed solutions involve aggregation features and many works can be seen as special cases [3, 18]. Our aim is to characterize each aggregation technique apart in order to determine its suitable use situations. We present in the following these different aggregation techniques

[1] http://www.naver.com

3.1 Aggregation by Clustering

Clustering is a technique that can be used to improve the user results space as it groups similar content based on topical coherence [3]. However, it is not enough to just return clusters of results [2]. It is important to provide users with an overview of the documents content forming each cluster to guide him in his search.

The utility of clustering in information retrieval is to enable the disambiguation and easiness of access to information [3]. Some studies claim that the grouped presentations were more effective than sequential presentations [11, 12]. While clustering is useful in information retrieval process for identifying relevant content, it is used to organize the content deemed estimated as relevant in aggregated search.

Clusty-yippy[2] is a search engine that queries other search engines and then aggregates the results together into clusters. Google news[3] is also an interesting example of aggregation by clustering. It aggregates news from various sources. The results are not a single list of news, but cluster of news. Each Cluster concerns a simple story. A short summary of one of the most representative item is given as Cluster title.

3.2 Multi-documents Summarization

Multi-document summarization has been used in *WebInEssence* [20] as a technique of aggregation. *WebInEssence* is a personalized Web-Based Multi-Document Summarization and recommendation system. It is designed to help users to find useful information in selected documents based on personal user's profiles. Results are presented in the way of automatic document summaries. This technique is interesting to alleviate the problem of information overload and to help more users to find the information they need. The summarization is the process of selecting the most significant information from a document. When the input consists of more than one document, we talk about multi-document summarization.

3.3 Document Generation

Document generation aggregation techniques seek to build or automatically generate a document from multiple documents of the same or different sources [14, 16 and 17].

Authors in [17] build automatically medical articles for Wikipedia from models they themselves generate. Content is then selected from Internet for each part of the model (eg. diagnosis, symptoms, causes…). Authors in [14, 16 and 26] used techniques from Natural Language Generation (NLG) [25] to create an aggregated result. NLG theories were used to determine the best organization of retrieved information. The organization of the returned information is defined by the role that each piece of information plays as well as the relationships between different pieces of information.

[2] http://yippy.com/.
[3] http://www.news.google.com

3.4 Relational Aggregation

Other works have exploited the structures contained in the Web as tables and lists [15]. In [18], the authors have developed an aggregation technique based on attributes retrieval. An aggregated tabular result of the form “attribute / value” is built for each query through three steps: The selection of entities and attributes relevant to the class designated by the query, the filtering of retrieved attributes and finally, the sorting of relevant attributes. In the same case, Google Labs[4] launched an experimental tool, called *Google Squared*, which generates a descriptive table for a given query. Figure 1 shows an example of Google Squared results for the query “cheeses”.

Fig. 1. Example of a Google Squared results for the query “cheeses”.

3.5 Aggregated Views

Previously mentioned aggregation techniques try to construct or generate a response (a document) from different sources. Aggregated views technique merges the different results of the sources in a single results page. This technique is used in web aggregated search. It is therefore question of presenting the search result page as an aggregated view of different search results. In this case, each information sources returns a single type of media (video, image, text...) or content (news, blog, news, products...). There are two types of aggregated views: the “blended view" and the “non-blended view” [24]. In the former one, heterogeneous results from different sources are merged and presented vertically in a single list. Google universal use this way to present the search result page. Other retrieval engines like Yahoo Alpha[5]and ASK3D[6] use a "non blended view" where each type of results is displayed in a separate part (panel) of the search results page.

[4] http://www.googlelabs.com
[5] au.alpha.yahoo.com/
[6] http://www.ask.com

4 Aggregation Techniques Study

4.1 Which Criteria?

Via, this study, we intend to find some criteria that allow the characterization of each aggregation technique. We have deduced that in each situation, some aggregation techniques are more suitable to use than others. To define these criteria, we should determine the main elements which can be involved in a multi-sources information retrieval environment. For our study, we think that the following criteria are the main criteria characterizing a use situation for an aggregation technique: The information need (the query), the user profile and the information sources content.

Firstly, the Query Itself or the Information Need. It is obvious that the kind of the information need greatly influences the choice of an aggregation technique. We will include a number of specific information needs that are associated with appropriate aggregation techniques.

For "named entities" queries, relational aggregation will directly stated and seems more suitable. We recall that relational aggregation technique tries to present the result as a descriptive table (attributes/values) for the entity designated by the query. A user who submits the query "Algeria", will certainly appreciates an answer as attribute / value where attributes are all attributes characterizing the entity class “Country” (Table 1). A user who submits the query "French presidents" will also appreciate an answer as a descriptive table containing a list of French presidents.

Table 1. Relational aggregation example.

Capital	Currency	President	Population	Off. Language
Algiers	Algerian Dinar	Abdelaziz Bouteflika	36M	Arabic

For vagueness (ambiguous) information needs, Clustering aggregation technique with multi-document summarization seems interesting. These techniques allow decreasing information overload and guide users to required information. They enable disambiguation and easiness of access to information by grouping results by concepts and giving an overview of each group contents via automatic summarization.

Some information need contain explicit intentions for a desired content type to retrieval. This is usually the case when terms like: *picture, image, photo, map, video, news* are present in the query string. Here aggregated views techniques (non-blended and blended views) have a great chance to satisfy the user. These techniques incorporate content other than standard web results as images, videos, blogs, and maps... in the results page. A query like « *new Renault CLIO pictures»* will be certainly satisfied by incorporating images results in the results page. Figure 2 shows results page for this query given by yahoo search engine. A query like « *Madonna last concert* » will certainly requires integration of some video clip to the results page.

Aggregated views will be also a suitable use situation for general information need (informational queries). With informational queries, users are seeking general information on a broad topic such as “natural phenomena” or “nutrition” [24]. In this case,

aggregated views allow presenting a rich and diverse response space which can concern several facets of the general information need.

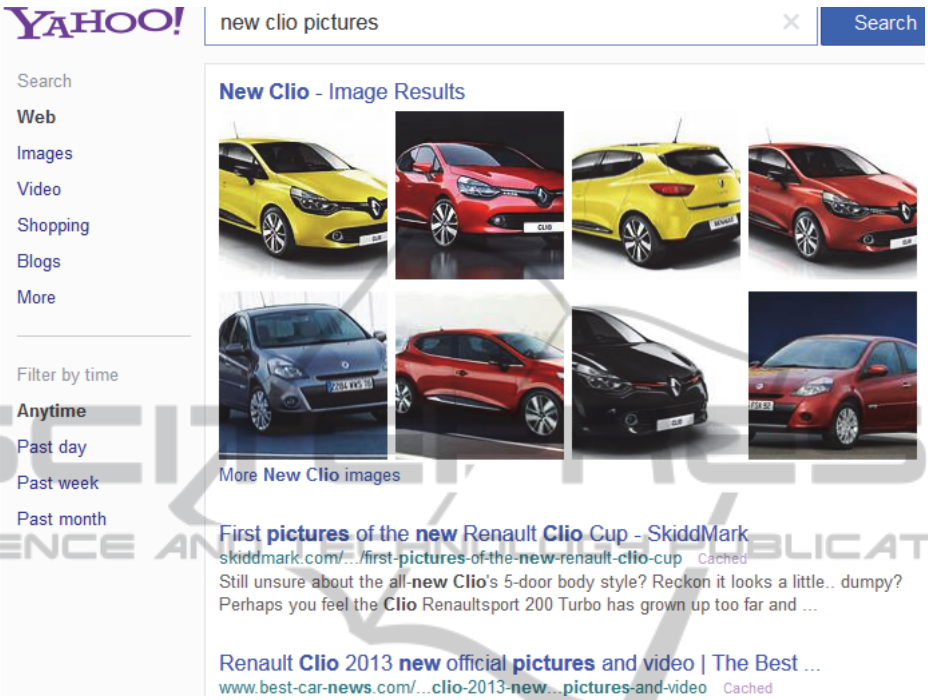

Fig. 2. Example of query with vertical intent images.

Secondly, the User Profile, more Precisely, the User Task. The user profile can make some aggregation technique more suitable than others. We associate here user profile to the user task. User retrieve information required for doing some tasks. Indeed, for some user task, the structure of the desired response can be known and thus modeled before searching. In these cases, the user query is considered as an input to a document generating system. For a user who is responsible for producing monitoring report, a user who is responsible for producing brochures or user who produce customized training guide for students .... For all these tasks the structure of the response is predictable at the beginning. For example, touristic brochure will contain a response composed by a presentation, some pictures, hotel information, and weather information...Document generation aggregation techniques are fully suitable in these scenarios. They try to generate coherent answers for repetitive queries. They try in first to define the structure of the response for the query. The response is then generated by retrieving information for each part of defined response.

Thirdly, Sources Content Type. We believe that the information sources content influence aggregation techniques. In the presence of heterogeneous (many modalities) content sources (text, images, video ...), aggregated view techniques (blended and non-blended) are more appropriate even if other techniques are not excluded. Cluster-

ing and summarization are typically used in environments of information sources of the same content type (text content). Aggregated views techniques allow representing heterogeneous results simultaneously in the same results page. The results are merged vertically or placed in different parts (panels) of the page. Each panel is designated to receive one results format.

4.2 Aggregation Techniques According to Suitable Use Situations

Table 2 summarizes the previous study and present suitable use situations for each aggregation technique.

Table 2. Mapping between aggregation techniques and use cases.

Aggregation technique	Criteria		
	Information need	**User profile(task)**	**Sources content**
Clustering	Vagueness		
Multi-documents summarization	Vagueness		Textual content
Relational aggregation	Named entity queries		
Document generation	Complex information need	Repetitive Task	
Aggregated views	Query with vertical intent. Informational queries.		Several modality content

The study shows that ach of the techniques can be suitable for a situation and less suitable for another situation. The finding states that aggregated search is a wide problem that cannot be solved or treated as a whole. Each aggregation technique was developed for a specific situation. We can see that some aggregation techniques have more large area of use than others. For example, document generation and relational aggregation techniques have a limited area of use. The former is strictly limited to a specific domain and very specific user profile (user task) for which the technique is developed. Relational aggregation is limited to named entity queries. Aggregated views and clustering techniques are less restrictive. They have more large area of use.

5 Conclusions

Aggregated search is a new multi-source information retrieval paradigm. It was never treated as a whole. Several developed solutions involving aggregation features and many works that can be considered as specific cases have been reported in literature and industry. In this paper, we described this information retrieval paradigm; we presented and studied, in term of use, aggregation techniques developed in this context.

The main contribution of the paper is an attempt to characterize the aggregation techniques developed in aggregated search according to usability perspective. We performed a description of these techniques and deduce some criteria that permit to feature suitable use situation for each of them. To our knowledge, this is the first at-

tempt to study the techniques from this point of view, i.e.; usability. We identified three main characterizing criteria that permit to distinguish between aggregation techniques reported in this paper: The information need (the query), the user profile (limited here to the user task) and information sources content.

The study of the aggregation techniques shows that each of the techniques can be suitable for a specific situation. This study confirms previously mentioned assertion, i.e.; aggregated search has never been treated as a whole [18]. Each aggregation technique was developed for a specific situation. Some techniques were developed for some kind of queries, some others were developed for a very specific user profile and others are most general and developed according to sources content types. Thus, it is important to select a context or an application case before thinking about the development or the use of an aggregation technique.

Undeniably, this study needs empirical experiments in order to confirm findings and conclusions. Other characterizing criteria such as those related to visualization constraints, semantic aspects and context might also be considered. We are interesting in future work to confirm the finding by empirical experiments and to develop dynamic aggregation technique.

References

1. V. Murdock and M. Lalmas, editors. SIGIR 2008 Workshop on Aggregated Search, New York, NY, USA, ACM. 1, 3.7. 2008.
2. S. Sushmita, M. Lalmas, A. Tombros . Using digest pages to increase user result space: Preliminary designs. Proceedings of the 2008 ACM SIGIR Workshop on Aggregated Search. Singapore, July 2008.

A. Kopliku, K., M. Boughanem. Aggregated search: potential, issues and évaluation. Technical report: IRIT/RT–2009-4–FR, IRIT, septembre 2009.

3. J. Callan, Distributed information retrieval. In W. Croft (Ed.), Advances in Information. Retrieval (pp. 127–150). Hingham, MA, USA: Kluwer Academic.2000.
4. J. Callan and M. Connell. Query-based sampling of text databases. ACM Transactions on Information Systems, 19:97–130, 2001.
5. R. Caruana. Multitask learning. Machine Learning, 28:41–75, July 1997.
6. J. Caverlee, L. Liu, and J. Bae. Distributed query sampling: A quality conscious approach. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’06, pages 340–347, New York, NY, USA, 2006.
7. K. Chen, R. Lu, C. K. Wong, G. Sun, L. Heck, and B. Tseng. Trada: Tree based ranking function adaptation. In Proceeding of the 17th ACM conference on Information and knowledge management, CIKM ’08, pages 1143–1152, New York, NY, USA, 2008..
8. L. Gravano, Chen-Chuan K. Chang, H. García-Molina, and A. Paepcke. Starts: Stanford proposal for internet meta-searching. In Proceedings of the 1997 ACM SIGMOD international conference on Management of data, SIGMOD ’97, pages 207–218, New York, NY, USA, 1997.
9. M. Shokouhi. Central-rank-based collection selection in uncooperative distributed information retrieval. In Proceedings of the 29th European conference on IR research, ECIR’07, pages 160–172, Berlin, Heidelberg, Springer-Verlag, 2007.
10. W. Rivadeneira and B. B. Bederson. A study of search result clustering interfaces: Comparing textual and zoomable user interfaces. Technical report, University of Maryland HCIL Technical Report HCIL-2003-36, 2003.

A. Becks, Ch. Seeling, and R. Minkenberg. Benefits of document maps for text access in knowledge management: a comparative study. In SAC ’02: Proceedings of the 2002 ACM symposium on Applied computing, pages 621–626, New York, NY, USA, 2002.
11. M. J. Cafarella, A.Y. Halevy, Y. Zhang, D. Zhe Wang, E. Wu., "Uncovering the Relational Web." In: Proceedings of WebDB (WebDB 2008), 2008.
12. C. Paris, N. Colineau, A. Lampert and K. Vander Linden. Discourse Planning for Information Composition and Delivery: A Reusable Platform. Journal of Natural Language Engineering, 16(1):61-98, Cambridge University Press, 2010.
13. H. Elmeleegy, J. Madhavan, A. Y. Halevy, Harvesting Relational Tables from Lists on the Web. Proceedings of the VLDB Endowment (PVLDB), 2009, pp. 1078-1089. 2009.
14. C. Paris, S. Wan, P. Thomas : Focused and aggregated search: a perspective from natural language generation. Information Retrieval, Vol. 13, issue 5, 434-459. October 2010.
15. C. Sauper, R. Barzilay, Automatically Generating Wikipedia Articles: A Structure-Aware Approach, In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 208-216, August 2009.
16. Krichen, A. Kopliku, K. Pinel-Sauvagnat, and M. Boughanem, Une approche de recherche d'attributs pertinents pour l'agrégation d'information. In Proceedings of INFORSID, 385-400, 2011.
17. M. Boughanem, J. Savoy., Recherche d'information états des lieux et perspectives, Hermès Science, Avril 2008.
18. D. R. Radev, W. Fan, and Z. Zhang. Webinessence: A personalized web-based multi-document summarization and recommendation system. In *NAACL Workshop on Automatic Summarization*, Pittsburgh, PA, 2001.
19. T. Truong Avrahami, L. Yau, L. Si, and J. Callan. The fedlemur project: Federated search in the real world. J. Am. Soc. Inf. Sci. Technol., 57(3):347–358, ISSN 1532-2882. 2006.
20. N. E. Craswell. Methods for Distributed Information Retrieval. PhD thesis, The Australian National University, vi, 29, 30, 31. May, 2000.
21. Arguillo. Federated Search for Heterogeneous Environments. PhD thesis, Carnegie Mellon University, CMU-LTI-11-008. 2011.
22. S. Sushmita. Study of results presentation and interaction for aggregated search. Phd Thesis, University of Glasgow. 2012.
23. W. Mann and S Thompson. Rhetorical structure theory: towards a functional theory of text organization. Text. 8(3), 243-281.1998.
24. C. Paris and N. Colineau. Scifly: Tailored corporate brochures on demand. . CSIRO ICT Centre Technical Report TR06/06. 2006.