

# From Aggravated to Aggregated Search: Improving Utility through Coherent Organisations of an Answer Space

Stephen Wan<sup>1</sup>      Cécile Paris<sup>1</sup>  
ICT Centre<sup>1</sup>  
CSIRO  
North Ryde (Sydney), Australia  
Firstname.Lastname@csiro.au

Alexander Krumpholz<sup>1,2</sup>  
Department of Computer Science<sup>2</sup>  
The Australian National University  
Canberra, Australia  
Krumpholz@acm.org

## ABSTRACT

Often, information seeking tasks require a number of related queries and rely on information from a number of diverse sources. Currently, the burden of issuing these individual queries, analysing for relevancy, as well as aggregating results falls upon the time-poor and informationally overloaded user. This paper argues that an appropriate organisation of the retrieved information helps the user to make sense of how each contributing piece of information from a variety of different and potentially heterogeneous sources relate to each other. We advocate the combination of methods from Natural Language Processing (specifically, text planning and summarisation research) and Information Retrieval for this purpose: namely, that of presenting search results using a coherent organisation and thereby ameliorating the tedium and aggravation involved in extensive queries. We demonstrate our hybrid approach, which is designed to capitalise on approaches that make use of data with both structured and unstructured components, with a system that enriches patient records with summaries of, and access to, relevant PubMed medical literature and its associated meta-data.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Information filtering, Search process*;  
H.5.2 [Information Interfaces and Presentation]: User Interfaces—*Natural language*; H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia—*User issues*

## General Terms

Design, Human Factors

## Keywords

Text Clustering, Automatic Text Summarisation, Text Planning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR'08, Workshop on Aggregated Search July 24, 2008, Singapore. The copyright stays with the authors.

Copyright 2008 ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

## 1. INTRODUCTION

We address the problem of supporting complex information seeking tasks which typically require a number of related queries and rely on information from a number of diverse sources, each with its own access interface. Currently, the burden of issuing these individual queries is typically on the users. The responsibility for analysing the results also falls on the users who need to identify the portions of returned documents and consolidate disparate pieces of information. The multiplicity of the various queries and the further identification of relevant extracts introduces a tedium endemic to complex tasks that runs the risk of aggravating the often time-poor user.

To address these needs and support the underlying task of the user, we explore mechanisms to automate parts of the process and streamline the retrieval of relevant information. Often the design of such a system considers implementation issues regarding the specification of related queries, utilisation of different information sources and identification of the appropriate granularity of information.

However, presentational issues are also pertinent. On any display medium, from computers and low-bandwidth handheld devices, and even to paper, organising the information is an important element to aggregated search. This paper argues that an appropriate organisation of the retrieved information helps the user to make sense of how each contributing piece of information from a variety of different sources relate to one another.

We introduce the concept of the “answer space”. This is a recognition that: (1) an information need might not be addressed by a single fact, web link or data source, but rather through the integration of multiple (potentially heterogeneous) sources, and (2) a single system might not be able to completely provide all the relevant information, or “the answer”. An answer space is a restricted set of data and data sources, arranged in a way that allows the reader to navigate through the space in a natural and effective way. It is worth noting that an answer space might include more than just the information deemed “on topic”. It might include supporting information and justifications for the content provided. This might help readers understand why some information is worth examining. It might also increase the users’ trust in the information (or allow them to decide whether the information is trustworthy, by considering the reliability of its source, for example).

The relationships underlying data elements in the answer space are complex and varied. They can be heterogeneous ranging from extracted text passages from multiple docu-

ments of differing domains and genres, database entries, or meta-data tagged graphics, to mention but a subset. Presenting these elements as a ranked list of separate items, as search engines conventionally do, still places the onus on the user to find a coherent organisation of the information. Instead, to arrive at a more meaningful and useful presentation of answers, the approach proposed in this paper considers how alternative organisations of the information might help the user make sense of the complex answer space.

There are a number of ways to determine an information organisation. It can be dynamically planned top-down by explicitly decomposing information needs into its components and imposing an organisation. Alternatively, it can be discovered in a bottom-up fashion by discovering and analysing relations already present in the data. Furthermore, an explicit representation of an organisation can be dynamically reasoned about or statically hard-coded.

In our previous work, top-down and bottom-up approaches, taken predominantly from the fields of Natural Language Processing (specifically, text planning and summarisation research) have been coupled with work in Information Retrieval. In this paper, we advocate the hybrid of the top-down and bottom-up approaches that is especially relevant for scenarios where textual data has both a structured and unstructured data components, for example meta-data associated with free text.

In the remainder of the paper, we highlight the main points of the two approaches and illustrate these with implemented systems. To better determine and organise the answer space, we describe our hybrid approach in terms of work which ties medical literature to a patient history. We argue that a hybrid approach should be used to leverage the individual strengths and offset the weaknesses of the separate approaches. To begin with however, we first examine related work to describe the linguistic foundations of structuring answer spaces.

## 2. RELATED WORK

This section first describes work on organising information from Natural Language Generation, providing a linguistically motivated argument for organising answer spaces. It then describes related work in Automatic Text Summarisation that attempts to answer similar questions to research in aggregated search.

### 2.1 Lessons from Natural Language Generation

The observation that information needs can be more complex, often requiring a complex answer is akin to the observation McKeown [7] made in her work on producing paragraph length text in the context of database queries. In that work, McKeown argued that some queries are such that people expect their answer to include several facts, organised in some prototypical fashion, referred to as a *schema*.

This organisation is reflected naturally in the way one writes textual documents. Because of the way people normally communicate, one tends to juxtapose one piece of information with another for deliberate reasons. That is, the organisation of a (formally) written document is usually not random. For example, to describe an object, one typically defines the object in terms of its superordinate(s) (e.g., *a stool is a piece of furniture*), provides its components and/or its attributes, compares it to an object probably known to

the user or differentiates it from another object which could easily be confused with the original object (e.g., *a stool is like a chair except that...*), and finally gives a concrete example of it.

As a result, in requesting a description of an object, one expects an answer comprising these facts, organised in that manner. From a natural language perspective, these typical ways of talking (writing) are referred to as “discourse strategies”. These are usually defined based on corpus analysis (done through a manual or statistical process).

Schemas, however, are static. Other approaches allow this organisation to be dynamic depending on the changing information need and the availability of related content, for example, see Moore and Paris [8]. Moore and Paris plan the document by decomposing the information need into component sub-goals. This decomposition is determined by a linguistic theory called Rhetorical Structure Theory (RST) [6] which describes the different relations one information element can have with another. Moore and Paris use this approach for planning and generating sentences for a system that plans and generates educational dialogue.

We base some of our work on notions taken from RST and the planning approaches in [8]. One disadvantage to the above text planning approaches, however, is its reliance on symbolic knowledge bases which are difficult to obtain. This limitation is typically referred to as the knowledge acquisition bottleneck. Nevertheless, the core linguistic ideas of coherent organisation are still useful and the systems described capitalise on these.

### 2.2 Text Analysis and Summarisation

Summarisation research is relevant to the goals of aggregated search in identifying relevant and important segments of a document for presentation to the user. Recently, research has moved from describing the important information in a single document to multiple related documents, for example see [2]. Query and Topic-based summarisation have also been explored and allow the summary to specifically target an information need. This has been the focus of shared-task workshops like Document Understanding Conference.<sup>1</sup>

Research in Automatic Text Summarisation, particularly those that produce summaries by extracting portions of text, sometimes draws on Information Retrieval approaches. For example, the use of vector space approaches to represent text portions, in this case sentences, and the use of the cosine similarity metric has been explored by a number of researchers, for example [11]. INEX<sup>2</sup> conferences have also seen the use of XML structures for part-of-document extraction.

Recent summarisation approaches tend to highlight relationships between sentences. These can be clusters of related sentences, again based on cosine similarity (for example, [11]), or more linguistically motivated features, as in the work on lexical chains by Barzilay [1]. These methods are related to text segmentation approaches (for example, [5] and more recently [3]) which break a text up into appropriately sized portions based on statistical patterns in word usage.

In essence, these approaches say something about the structural relationships within the text based on content, though they are not as fine-grained as the organisation derived from

<sup>1</sup><http://duc.nist.gov/>

<sup>2</sup><http://www.inex.otago.ac.nz/>



Figure 1: A dynamically generated SciFly PDF brochure

RST. Although more robust than symbolic generation approaches, extraction-based summarisation approaches may be less precise due to errors introduced in its use of text analysis methods.

### 3. ORGANISING THE ANSWER SPACE

We illustrate each of the strengths and weaknesses of the top-down and bottom-up approach with systems developed for various aggregated search scenarios. Each of these combines the methods from Information Retrieval, Text Planning, and Text Summarisation to address multiple aspects of aggregated search.

#### 3.1 Driving Retrieval with Coherent Organisations

The top-down approach begins by first planning a generated document organisation that meets the user’s information need. This is then populated with data returned via search engines. We demonstrate this approach on a personalised brochure generation system.

We plan a possible document organisation by beginning with the goal of the document, that is, the user’s information need. This goal decomposes into sub-goals, leading eventually to sub-queries, which must be met for the document to be planned successfully. Crucially, each of these sub-goals has constraints on the coherent relationship between them. This produces a tree structure where sibling nodes have explicit annotations encoding the representing the rhetorical relationship between the siblings’ information.

For example, in an information system in the tourism domain[10], tourist information about some city must be supported by information that *enables* the user to get there in the first place, say by presenting transportation options (which further decomposes into sub-goals). This relationship is an example of an *enablement* relationship.

Instead of relying on hand-crafted knowledge bases, one can couple text planning approaches with information retrieval methods. The aim here is to re-use text from passages retrieved by a search engine instead of generating it from scratch.

The leaf level of the tree — those without further decomposition — has sufficient context to issue a query. At this point, personal information about the user is also au-

tomatically included to tailor the search. Using such a text planning mechanism, the results of the search for the particular sub-query already has an appropriate location into which one can slot the results. We show in [10] that driving the retrieval process using text planning leads to organised generated documents that were favoured by participants.

This approach has been implemented as a domain-independent system and trialed on the tourism domain and on an enterprise document set [9]. In the latter case, the system, SciFly, creates a personalised brochure which describes information about an organisation, in this case Australia’s science organisation CSIRO. An example is presented in Figure 1 which shows a brochure that repurposes marketing and public relations data retrieved to meet a user’s query. One limitation to the work is that it currently requires structured data sources that constrain the content types returned.

### 3.2 Finding Data Relationships in Text

In contrast, a bottom-up approach examines the text to see what supplementary information *already exists* in the text(s) returned by a search engine. Specifically, we organise information by investigating the role of automatically discovered supplementary context in understanding extracted sentences.

Our preliminary investigations on this work, referred to as *Elaborative Summarisation*, are described in [12] which presents sentence extraction summaries of linked documents using the reading context of user as an means to personalise the summary. In the domain of Wikipedia<sup>3</sup> text, a summary of a linked document that is as yet unseen by the user is generated in order to find information that *expands* on the content of the linking sentence. In short, we use the reading context to define the information need.

The basic approach determines the relatedness of sentences in the *linked* document to clusters of sentences in the reading context, in particular the cluster or focus represented by the linking sentence. Sentence relatedness is currently obtained using Singular Value Decomposition following [4]. The system identifies the *neighbourhood*, a text passage, in which elaborative material within the linked document may be found. In a preliminary evaluation, recall and precision scores above a baseline are achieved.

### 4. A HYBRID APPROACH

Each approach has its strengths and weaknesses. In ongoing work, we explore further possibilities in combining top-down and bottom-up techniques in order to create coherent answer spaces from documents with both structured and unstructured data. We investigate this in the context of a project that enriches timeline views of medical histories with references to medical literature. The aim is to support a doctor’s decision-making processes by retrieving information related to a patients history from the PubMed publication archives.<sup>4</sup>

PubMed currently only provides meta information like publication dates, the journal title, the authors, the title, the abstract and relevant key terms. The key terms come from a set of medical subject headings<sup>5</sup> (MeSH) and provide content-based descriptions on the paper.

<sup>3</sup><http://en.wikipedia.org/>

<sup>4</sup><http://www.pubmed.gov/>

<sup>5</sup><http://www.nlm.nih.gov/mesh/>

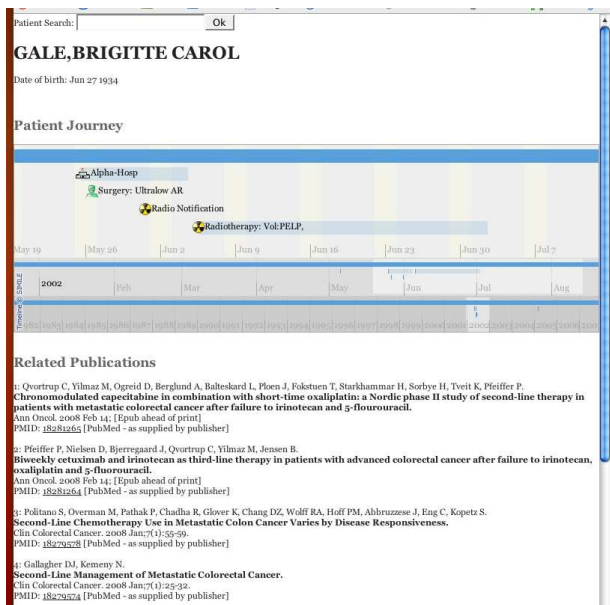


Figure 2: Enriched Patient History Timeline

The basic query is made up of terms extracted from the patient history which includes key words and events. Key words like “cancer” or “stage” are used as query terms. Events such as the patient’s date of birth, hospital submission date, diagnosis date and treatment dates are used to filter relevant documents. These are then used to populate a query and retrieve relevant PubMed documents. The goal is to present the highly ranked documents along with a timeline view of the patient history.

In this multi-document summarisation scenario, bottom-up clustering approaches based on the contents of each document and the MeSH terms are used to discover relationships between the documents. To present the results in a coherent fashion, a summary for a cluster is generated using the the MeSH meta-data intersection of documents in that cluster. Text planning approaches describe how documents and information in a cluster relate to one another.

For a query containing the search phrase “colorectal cancer” for example, a schema for drill-down navigation might suggest that papers may be described in two groups depending on whether the experimental subjects are animals or humans. In this way, salient contrasting features may be used to show the relationships — in this case, a *contrast* relationship — between the documents in the answer space.

Sentence-extraction summarisation approaches can then be used to find the most indicative text in the abstracts to represent the cluster. If possible, further elaborative material is presented. In this way, a doctor viewing a patient’s history is able to see relevant medical literature aligned with events in the history and is able to see how the publications relate to one another in order to quickly determine relevance. Figure 2 presents an early screenshot of the enriched timeline based on a mock history.

## 5. CONCLUSION

In this paper, we advocate a hybrid approach of top-down discourse planning approaches with bottom-up text sum-

marisation (including text clustering) approaches to organise the answer space of an aggregated search scenario for use with text with associated meta-data. This was illustrated with a description of ongoing work in the area of medical publications that combines approaches in order to facilitate a doctor’s treatment of a patient by providing relevant literature and outlining the relationships between each documents.

## 6. REFERENCES

- [1] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, 1997.
- [2] R. Barzilay and K. R. McKeown. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328, 2005.
- [3] F. Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 26–33, Seattle, USA., April 2000.
- [4] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25. ACM Press, 2001.
- [5] M. Hearst. Multi-paragraph segmentation of expository text. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 9 – 16, New Mexico State University, Las Cruces, New Mexico, 1994.
- [6] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.
- [7] K. R. McKeown. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, 1985.
- [8] J. Moore and C. Paris. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Journal of Computational Linguistics*, 19(4):651 – 694, December 1993.
- [9] C. Paris and N. Colineau. Scifly: Tailored corporate brochures on demand. Technical Report 06/268, CSIRO ICT Centre, 2006.
- [10] C. Paris, S. Wan, R. Wilkinson, and M. Wu. Generating personal travel guides - and who wants them? In *Proceedings of the International Conference on User Modelling (UM2001)*, Sonthofen, Germany, uly 13-18 2001.
- [11] D. Radev, J. Otterbacher, and D. T. H. Qi. Mead reduces: Michigan at duc 2003. In *Document Understanding Conference 2003: Workshop on Text Summarization*, Edmonton, Canada, May 2003.
- [12] S. Wan and C. Paris. In-browser summarisation: Generating elaborative summaries biased towards the reading context. In *The 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Paper*, Columbus, Ohio, June 2008.