Multi-document Summarization by Visualizing Topical Content

Rie Kubota Ando
Department of Computer Science, Cornell University, Ithaca, NY 14853-7501
kubotar@cs.cornell.edu

Branimir K. Boguraev and **Roy J. Byrd** and **Mary S. Neff**
IBM T.J. Watson Research Center, 30 Saw Mill River Road, Hawthorne, NY 10532
{bran, roybyrd, maryneff}@watson.ibm.com

Abstract

This paper describes a framework for multi-document summarization which combines three premises: coherent themes can be identified reliably; highly representative themes, running across subsets of the document collection, can function as multi-document summary surrogates; and effective end-use of such themes should be facilitated by a visualization environment which clarifies the relationship between themes and documents. We present algorithms that formalize our framework, describe an implementation, and demonstrate a prototype system and interface.

1 Introduction: multi-document summarization as an enabling technology for IR

The rapid growth of electronic documents has created a great demand for a navigation tool to traverse a large corpus. Information retrieval (IR) technologies allow us to access the documents presumably matching our interests. However, a traditional hit list-based architecture, which returns linearly organized single document summaries, no longer suffices, given the size of a typical hit list (e.g. submitting the query "summarization workshop" to a search engine Altavista (`http://altavista.com`) gave us more than ten million hits).

To allow a more comprehensive and screen space-efficient presentation of query results, we propose in this paper a technology for summarizing collections of multiple documents. In our work, we focus on identifying themes, representative of a document, and possibly running across documents. Even if we are unable to 'embody' a theme in coherently generated prose, we start with the assumption that a mapping exists between a theme and a tightly connected (and therefore intuitively interpretable) set of coherent linguistic objects, which would act as a 'prompting' device when presented to the user in an appropriate context. As will become clear in the rest of the paper, we refer to such themes as *topics*.

Our view of multi-document summarization combines three premises: coherent topics can be identified reliably; highly representative topics, running across subsets of the document collection, can function as multi-document summary surrogates; and effective end-use of such topics should be facilitated by a visualization environment which clarifies the relationship between topics and documents. The work specifically addresses the following considerations.

• **Multiple general topics** We regard the ability to respond to multiple topics in a document collection — in contrast to a prevailing trend in multi-document summarization, seeking to present the single, possibly pre-determined, topic (see below) — to be crucial to applications such as summarization of query results. In this work we choose not to narrow the topic detection process by the given query, since in IR it is a well-known concern that user-specified queries do not necessarily convey the user's real interests thoroughly. Thus, we need to deal with multiple general topics.

• **Textual and graphical presentation** Since our multi-document summaries will, by definition, incorporate multiple topics, the question arises of optimal representation of the relationships among the topics, the linguistic objects comprising each topic, and the documents associated with (possibly more than one) topic. In particular, for IR, we want to show the relationships between topics and documents so that a user can access documents in the context of the topics. A topic by itself can clearly be represented largely by a set of text objects. However, we need also to present arbitrary number of such topics as part of the same summary. We believe that, for adequate representation of

the resulting many-to-many relationships (which is crucial for the end-user fully understanding the summary), additional graphical components are needed in the interface.

To our knowledge, the existing studies of multi-document summarization do not place emphasis on these considerations. Radev and McKeown (1998) have shown a methodology for 'briefing' news articles reporting the same event. Barzilay et al. (1999) have proposed a method for summarizing "news articles presenting different descriptions of the same event". These studies focus on a single topic in a document collection. Mani and Bloedorn (1999) have addressed summarizing of similarities and differences among related documents with respect to a specified query or profile. In their study, several presentation strategies are suggested. Although they mention a graphical strategy, such as plotting documents sharing more terms closer together, no implementation is reported.

There are a number of different studies that address graphical presentation of multi-document (or document corpus visualization) – The VIBE System (Olsen et al., 1993; Korfhage and Olsen, 1995), Galaxy (Rennison, 1994), SPIRE Themescapes (Wise et al., 1995), LyberWorld (Hemmje et al., 1994), and applications of self-organizing map utilizing neural network technique (Kohonen, 1997; Lin, 1993; Lagus et al., 1996). In general, these studies consider documents as objects in a model space (document space, typically high-dimensional) and provide 2-D or 3-D representation of this document space. Their focus is on detecting and presenting structural relationships among documents in a corpus.

From our viewpoint, these two fields of research address two different perspectives on the multi-document analysis problem: multi-document summarization efforts largely deliver their results in textual form, while document corpus visualization research, which focuses on means for graphical representation of a document space, does not perform any summarization work. While we believe that both textual and graphical representations are essential in the context of IR, the technologies from the two fields, in general, cannot be easily combined because of methodological differences (such as differences in modeling the document set, calculating similarity measures, and choosing linguistic objects in terms of which a summary would be constructed).

Motivated by these observations, we propose one uniform framework that provides both textual and graphical representations of a document collection. In this framework, topics underlying a document collection are identified, and described by means of linguistic objects in the collection. Relationships, typically many-to-many, among documents and topics are graphically presented, together with the topic descriptions, by means of a graphical user interface specifically designed for this purpose. We focus on relatively small document collections (e.g. 100 or so top-ranked documents), observing that in a realistic environment users will not look much beyond such a cut-off point. Our approach maps linguistic objects onto a multi-dimensional space (called *semantic space*). As we will see below, the mapping is defined in a way that allows for topics with certain properties to be derived and for linguistic objects at any granularity to be compared as semantic concepts.

The rest of this paper is organized as follows. The next section describes the multi-dimensional space for the document collection. Section 3 demonstrates our prototype system and illustrates the interplay between textual and graphical aspects of the multi-document summary. Section 4 highlights the implementation of the prototype system. We will conclude in Section 5.

2 Mapping a document collection into semantic space

Semantic space is derived on the basis of analyzing relationships among linguistic objects — such as terms, sentences, and documents — in the entire collection. A *term* can be simply a 'content word', in the traditional IR sense, or it can also be construed as a phrasal unit, further representative of a concept in the document domain. In our implementation, we do, in fact, take that broader definition of terms, to incorporate all types of non-stop lexical items as well as phrasal units such as named entities, technical terminology, and other multi-word constructions (see Section 4 below).

We map linguistic objects (such as terms, sentences, and documents) to vectors in a multi-dimensional space. We construct this space so that the vectors for the objects behaving statistically similarly (and therefore presumed to be semantically similar) point in similar directions. The vectors are called *document vectors*, *sentence vectors*, and *term vectors*, according to the original linguistic

objects they are derived from; however, all vectors hold the same status in the sense that they represent some concepts. In this work, we call this multi-dimensional space *semantic space* (Ando, 2000) to distinguish it from a traditional vector space (Salton and McGill, 1983). In essence, in our semantic space, the terms related to each other are mapped to the vectors having similar directions, while a traditional vector space model treats all terms as independent from each other.

Our motivation for using semantic space is at least twofold. First, we believe that we need the high representational power of a multi-dimensional space since natural language objects are intrinsically complicated, as Deerwester et al. (1990) argued. Secondly, our definition of semantic space allows us to measure similarities among concepts and linguistic units at any granularity. Single-word terms, multi-word terms, sentences, and topics – all can be equally treated as objects representing some concept(s) when they are mapped to vectors in this space. From the viewpoint of a summarization task, this is an advantage over a traditional vector space in which terms are assumed to be independent of one another.

To detect topics underlying the document collection, we create a set of vectors in the semantic space so that every document vector is represented by (or close to) at least one vector (called *topic vector*). In other words, we provide viewpoints in the semantic space so that every document can be viewed somewhat closely from some viewpoint. Given such vector representations for topics, we can quantitatively measure the degree of associations between topics and linguistic objects by using a standard cosine similarity measure between topic vectors and linguistic object vectors. The linguistic objects with the strongest association would represent the topic most appropriately.

The algorithm we use for semantic space construction (see Figure 5 in Section 4) is closely related to singular value decomposition (SVD) used in Latent Semantic Indexing (LSI) (Deerwester et al., 1990). As in SVD, this algorithm finds statistical relationships between documents and terms by computing eigenvectors, and it performs dimensional reduction that results in a better statistical modeling. The advantages of the semantic space we described above are shared with similar approaches (such as SVD-based and Riemannian SVD-based (Jiang and Berry, 1998)). The algorithm we adopt, however, differs from others in that it achieves a high precision of similarity measurement among *all* the documents by capturing information more *evenly from every document* while, with other approaches, the documents whose statistical behaviors are different from the others tend to be less well represented. This algorithm fits well in our framework since we want to find topics by referring the similarities of *all* pairs of documents (shown later), and also we want to assume *all the documents are equal*. Full details of the semantic space construction algorithm may be found in (Ando, 2000), including evaluation results compared with SVD.

3 Visual presentation of a semantic space: combining text and graphics

In this section, to illustrate how we combine textual and graphical presentation, we demonstrate a summary that our prototype system created from 50 documents (TREC documents relevant to 'non-proliferation treaty').

The document set is presented in one full screen in relation to the underlying topics. The prototype system detected six[1] topics in this document set (see Figure 1). For each topic, three types of information are presented: a list of terms (*topic terms*), a list of sentences (*topic sentences*), and a visual representation of relevance of each document to the topic (*document map*).

Below we highlight some essential features of the interface.

Topic terms and topic sentences: The topic presented at the upper right corner of Figure 1 has the topic terms "Iraq", "Iraqi", "Kuwait", "Saddam Hussein", "embargo", "invasion", "disarm", and so on. (The frame is scrollable, thus accommodating all topic terms.) A topic typically will be addressed by more than one sentence, presented in a closely associated scrollable frame. The first topic sentence for this topic is "Israel's Air Force bombed Iraq's Osirak ...". Together, the sets of topic terms and sentences describe the topic, i.e. one 'thread' discussed in possibly several documents.

Document proxy – a "dot" represents a document: In a document map, a dot image represents each document (i.e. *document proxy*). A dot before a topic sentence is also a document proxy representing the document containing that sentence.

[1]The number of topics detected depends on the document set and the parameter setting adjusting the granularity.

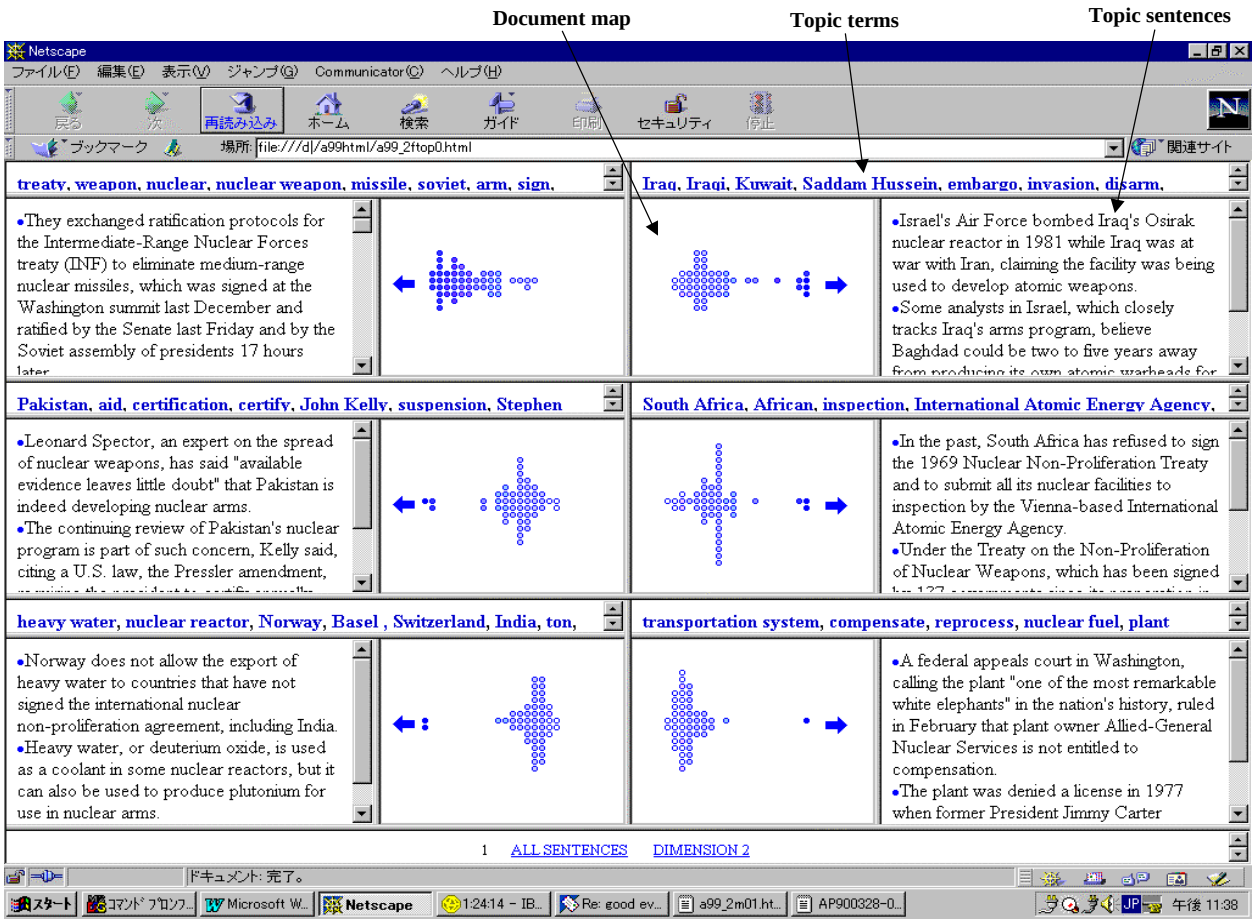

Figure 1: Example of the final output.

Document maps – topic-document relevance shown by document proxy placement and color gradation: In a document map, the horizontal placement of each dot represents the degree of relevance of the corresponding document to the topic. Documents closer to the direction of the arrow are more relevant to the topic. The color intensity of the dot also represents the degree of relevance. For instance, in the document map at the upper right corner of Figure 1, we see that there are six documents closely related to this 'Iraq-topic'. These six dots are placed on the right (the direction of the arrow), and their colors are more intense than the other document proxies. We see one more document to the left to the six documents, also with a relatively strong connection to this topic. Two documents, represented by dots almost at the center of the map, are only somewhat related to this topic. The rest of the documents, having dots that are almost transparent and placed on the left, are not very related to this topic. Thus, users can tell, at a glance, how many documents are related to each topic and how strongly they are related. Note that each document map contains proxies for all the documents. Unlike a typical clustering approach, we do not divide documents into groups. Clusters of documents, if any, are naturally observed in the document map. A document map is a projection of document vectors onto a topic vector. The semantic space allows us to detect and straightforwardly present the structural relationships among the documents.

Highlighting of document proxies – the relationships between a document and multiple topics: When a mouse rolls over a dot, the title of the document appears, and the color of the dots representing the same document in all the document maps changes (from blue to red) (see Figure 2). This color change facilitates understanding the relationships between a document and multiple topics.

A hot-link from a document proxy to full text:

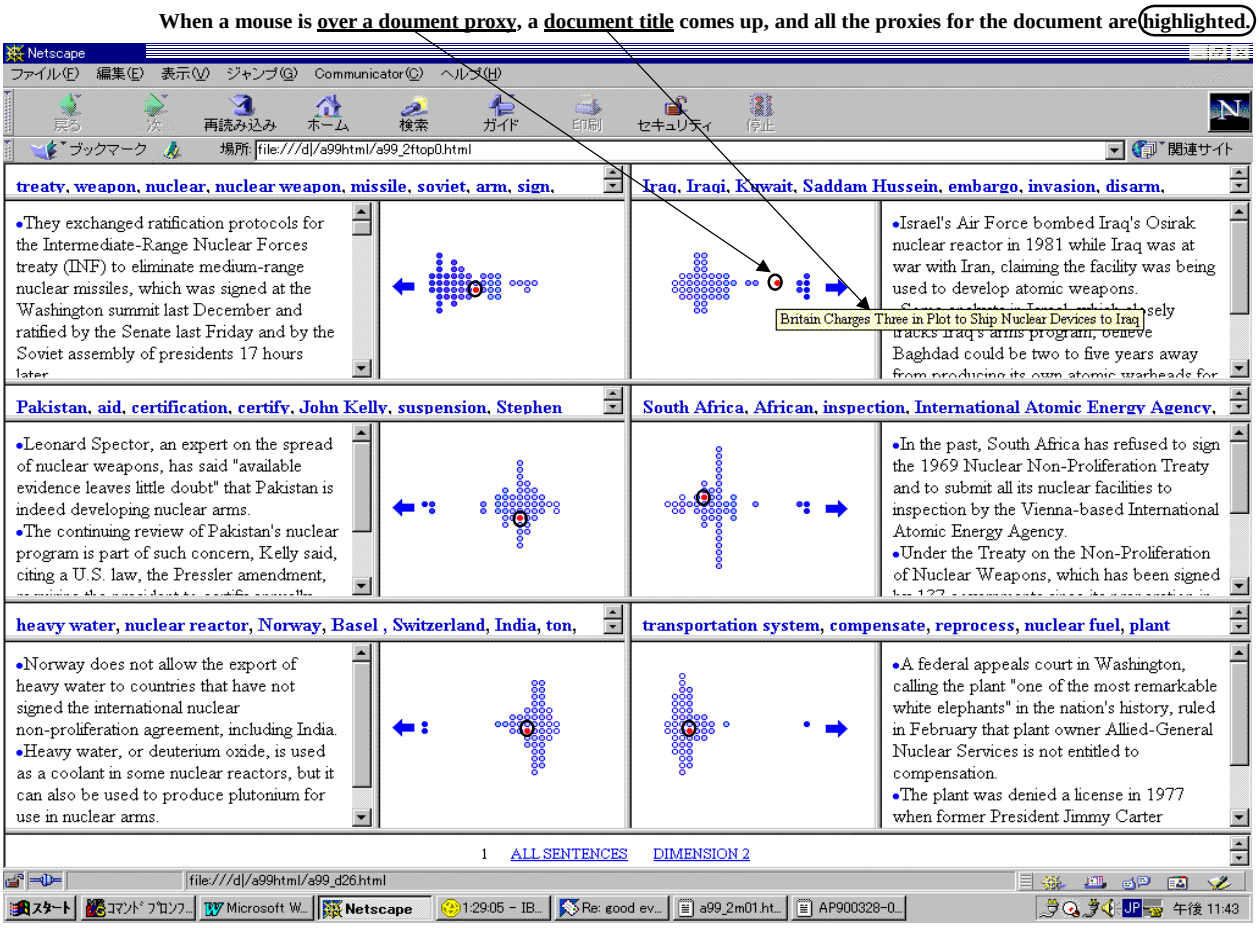

Figure 2: When a mouse rolls over a dot:

When a dot is clicked, the full text of the corresponding document is displayed in a separate window. This allows us to browse documents in the context of document-topic relationships.

Highlighting a topic sentence in the full text: When the clicked dot is associated with a topic sentence, the full text is displayed in a separate window, with the topic sentence highlighted. This highlighting helps the user to understand the context of the sentence quickly, and thus further facilitates focusing on the information of particular interest.

Topic sentences: Finally, we illustrate some of the topic sentences extracted by our system below. For each topic, the two sentences related to the topic most closely are shown.

'Iraq-topic':

• *Israel's Air Force bombed Iraq's Osirak nuclear reactor in 1981 while Iraq was at war with Iran, claiming the facility was being used to develop atomic weapons.*

• *Some analysts in Israel, which closely tracks Iraq's arms program, believe Baghdad could be two to five years away from producing its own atomic warheads for missiles or nuclear bombs to be dropped from jets.*

'Pakistan-topic':

• *Leonard Spector, an expert on the spread of nuclear weapons, has said "available evidence leaves little doubt" that Pakistan is indeed developing nuclear arms.*

• *The continuing review of Pakistan's nuclear program is part of such concern, Kelly said, citing a U.S. law, the Pressler amendment, requiring the president to certify annually that Pakistan does not possess a nuclear weapon.*

'South Africa-topic':

• *In the past, South Africa has refused to sign the 1969 Nuclear Non-Proliferation Treaty and to submit all its nuclear facilities to inspection by the Vienna-based International Atomic Energy Agency.*

• *Under the Treaty on the Non-Proliferation of Nuclear*

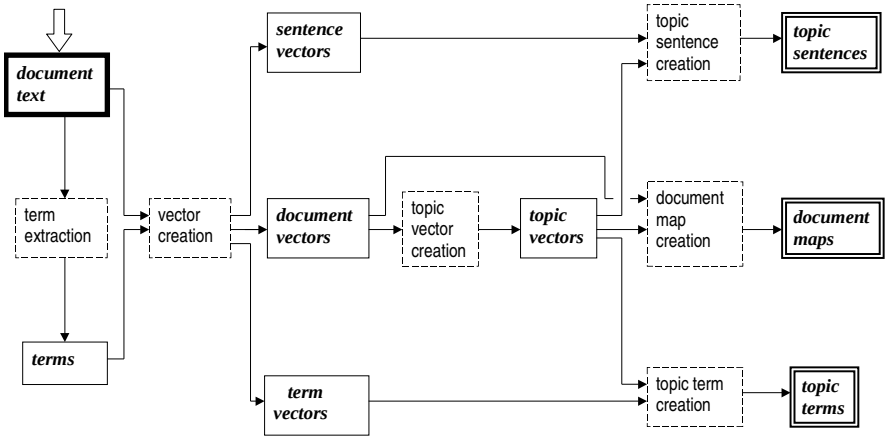

Figure 3: Overview of the process.
A block arrow indicates the input to the process, and rectangles with double-line border are the output. Rectangles with dashed line border are sub-processes. Other rectangles represent data.

Weapons, which has been signed by 137 governments since its preparation in 1969, countries without such weapons open their nuclear facilities to inspection by experts from the International Atomic Energy Agency, a U.N. agency based in Vienna.

Both for 'Iraq-' and 'Pakistan-topic', the two topic sentences address two different aspects of the similar "doubt" or "concern". For 'South Africa-topic', the second topic sentence gives background knowledge of the specific fact described in the first topic sentence. We find it interesting that, despite the fact that the two topic sentences are extracted from different documents, they appear to be consecutive sequences from a uniform source.

In essence, the design seeks to facilitate quick appreciation of the contents of a document space by supporting browsing through a document collection with easily switching between different views: topic highlights (terms), topical sentences, full document text, and inter-document relationships. At present, there is no attempt to handle redundancy between topic sentences.

4 Implementation

In this section, we describe the implementation of our prototype system. The overall process flow of this system is shown in Figure 3. Our description omits the process of creating graphical presentation that is straightforwardly understood from Section 3. The system takes, as its input, the text of a given set of documents. Throughout this section, we use the three small 'documents' shown below as an illustrative example. The data flow from these three documents to the final output is shown in Figure 4.

Document #1:
Mary Jones has a little lamb. The lamb is her good buddy.
Document #2:
Mary Jones is a veterinarian for ABC University.
ABC University has many lambs.
Document #3:
Mike Smith is a programmer for XYZ Corporation.

4.1 Term extraction

First, we extract all terms contained in the documents, using an infrastructure for document processing and analysis, comprising a number of interconnected, and mutually enabling, linguistic filters; which operates without any reference to a predefined domain. The whole infrastructure (hereafter referred to as TEXTRACT) is designed from the ground up to perform a variety of linguistic feature extraction functions, ranging from straightforward, single pass, tokenization, lexical look-up and morphological analysis, to complex aggregation of representative (salient) phrasal units across large multi-document collections (Boguraev and Neff, 2000). TEXTRACT combines functions for linguistic analysis, filtering, and normalization; these focus on morphological processing, named entity identification, technical terminology extraction, and other multi-word phrasal analysis; and are further enhanced by cross-document aggregation, resulting in some normalization to canonical forms, and simple types of

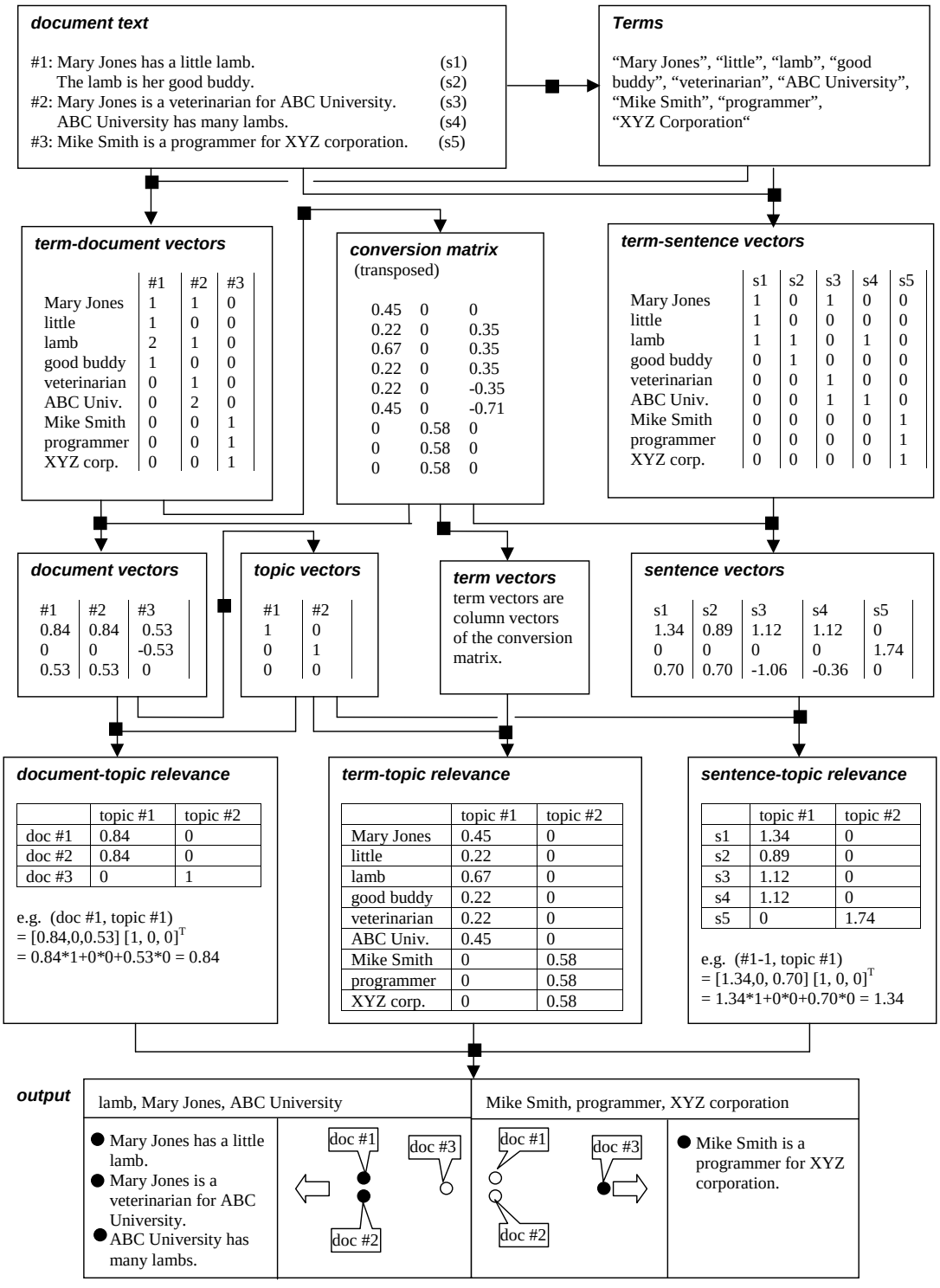

Figure 4: Example of data flow.

Procedure ConstructSemanticSpace
Input: term-document vectors $\mathbf{d}_1, ..., \mathbf{d}_n$
Output: conversion matrix $\mathbf{C}$

$\mathbf{D} = [\mathbf{d}_1 ... \mathbf{d}_n]$ /* Term-document matrix */
$\mathbf{R} = \mathbf{D}$ /* Initialize a residual matrix with the term-document matrix */
For $i = 1$ to k
  $\mathbf{R}_s = [|\mathbf{r}_1|^q \mathbf{r}_1 ... |\mathbf{r}_n|^q \mathbf{r}_n]$ /* Scale each of $\mathbf{R}$'s column vectors by a power of its own length */
  $\mathbf{c}_i =$ the eigenvector of $\mathbf{R}_s {\mathbf{R}_s}^T$ with the largest eigenvalue
  $\mathbf{R} = \mathbf{R} - [(\mathbf{c}_i{}^T \mathbf{r}_1)\mathbf{c}_i ... (\mathbf{c}_i{}^T \mathbf{r}_n)\mathbf{c}_i]$ /* Eliminate the direction of $\mathbf{c}_i$ from $\mathbf{R}$'s column vectors */
End for

$\mathbf{C} = [\mathbf{c}_1 ... \mathbf{c}_k]^T$ /* Conversion matrix */

Figure 5: Semantic space creation. Scaling factor q and the dimensionality k are experimentally determined.

co-reference resolution.

For the example mini-documents above, after removal of common stop words, the terms remaining as linguistic objects for the algorithm to operate on are listed at top of Figure 4.

4.2 Vector creation

We construct the semantic space from term-document relationships by a procedure[2] shown in Figure 5. In the semantic space, each of vector elements represents a linear combination of terms. The conversion matrix returned by the semantic space creation procedure keeps the information of these linear combinations. For instance, the conversion matrix for our example (see Figure 4) shows that the first element of a vector in the semantic space is associated with $0.45*$"Mary Jones"$+0.22*$"little"$+0.67*$"lamb"$+0.22*$"good buddy"$+0.22*$"veterinarian"$+0.45*$"ABC University".

To map the documents to the vectors in the semantic space, we create the *term-document vectors* each of whose elements represents the degree of relevance of each term to the document. Our implementation uses term frequency as the degree of relevance. We create document vectors of the semantic space by multiplying term-document vectors and the conversion matrix. Sentences and terms can also be mapped to the vectors in the same way by treating them as "small documents".

[2]We do not describe the details of this procedure in this paper. See Section 2.

4.3 Identifying topics

Ultimately, our multi-document summaries rely crucially on identifying topics representing all the documents in the set. This is done by creating topic vectors so that *each document vector is close to (i.e. represented by) at least one topic vector*. We implement this topic vector creation process as follows. First, we create a document graph from the document vectors. In the document graph, each node represents a document vector, and two nodes have an edge between them if and only if the similarity between the two document vectors is above a threshold. Next, we detect the connected components in the document graph, and we create the topic vectors from each connected component by applying the procedure 'DetectTopic' (Figure 6) recursively.

'DetectTopic' works as follows. The unit eigenvector of a covariance matrix of the document vectors in a set S is computed as $\mathbf{v}$. It is a representative direction of the document vectors in S. If the similarity between $\mathbf{v}$ and any document vector in S is below a threshold, then S is divided into two sets S_1 and S_2 (as in Figure 7), and the procedure is called for S_1 and S_2 recursively. Otherwise, $\mathbf{v}$ is returned as a topic vector. The granularity of topic detection can be adjusted by the setting of threshold parameters.

Note that such a topic vector creation procedure essentially detects "cluster centroids" of document vectors (not sentence vectors), although grouping documents into clusters is not our purpose. This indicates that general vector-based clustering technologies could be integrated into our framework if

it brings further improvement.

4.4 Associations between topics and linguistic objects

The associations between topics and linguistic objects (documents, sentences, and terms) are measured by computing the cosine (similarity measurement) between the topic vectors and linguistic object vectors. The degree of association between topics and documents is used to create document maps. The terms and sentences with the strongest associations are chosen to be the topic terms and the topic sentences, respectively.

As a result, for our example we get the output shown at the bottom of Figure 4.

4.5 Computational complexity

Let m be the number of different terms in the document set (typically around 5000), and let n be the number of documents (typically 50 to 100)[3]. Given that $m > n$, the semantic space is constructed in $O(mn^2)$ time. The topic vectors are created in $O(n^3)$ time by using a separator tree for the computation of all-pairs minimum cut[4], assuming that the document vector set is divided evenly[5]. Let k be the dimensionality of the semantic space, and let h be the number of detected topics. Note that k and h are at most n, but are generally much smaller than n in practice. Regarding the number of terms contained in one sentence as a constant, topic sentences are extracted in $O(skh)$ time where s is the total number of sentences in the document set. Topic terms are extracted in $O(mkh)$ time. We note that the prototype system runs efficiently enough for an interactive system.

5 Conclusion and further work

This paper proposes a framework for multiple document summarization that leverages graphical elements to present a summary as a 'constellation' of topical highlights. In this framework, we detect topics underlying a given document collection, and we describe the topics by extracting related terms and sentences from the document text. Relationships among topics and documents are graphically presented using gradation of color and placement of image objects. We illustrate interactions with

[3]In this work, we focus on relatively small document collections; see Section 1.

[4]See (Ahuja et al., 1993) for all-pairs min cut problem.

[5]Note that Step 3 in the document vector division procedure (Figure 7) seeks for this.

Procedure DetectTopic(S)
Input: a set of document vectors S
Output: topic vectors

v = the unit eigenvector of a covariance matrix of document vectors in S
Loop for each document vector **d** in S
  if similarity between **d** and **v** is below a threshold then begin
    divide S into S_1 and S_2
    Call DetectTopic(S_1)
    Call DetectTopic(S_2)
    Exit the procedure
  End if
End loop
Return **v** as a topic vector

Figure 6: Topic vector creation.

our prototype system, and describe its implementation. We re-emphasize that the framework presented here derives its strength in equal part from two components: the results of topical analysis of the document collection are displayed by means of a multi-perspective graphical interface specifically designed to highlight this analysis. Within such a philosophy for multi-document summarization, sub-components of the analysis technology can be modularly swapped in and replaced, without contradicting the overall approach.

The algorithms and subsystems comprising the document collection analysis component have been implemented and are fully operational. The paper described one possible interface, focusing on certain visual metaphors for highlighting collection topics. As this is work in progress, we plan to experiment with alternative presentation metaphors. We plan to carry out user studies, to evaluate the interface in general, and to determine optimal features, best suited to representing our linguistic object analysis and supporting navigation through query results.

Other future work will focus on determining the effects of analyzing linguistic objects to different level of granularity on the overall results. Questions to consider here, for instance, would be: what is the optimal definition of a term for this application; does it make sense to include larger phrasal units in the semantic space; or do operations over sentences, such as sentence merging or reduction, offer alternative ways of visualizing topical content.

Step1: Create a sub-graph G_s for S from the document graph.
Step2: Compute the minimum cut of all the node pairs in G_s.
Step3: Evaluate each minimum cut, and choose the cut (A, B) maximizing $h(A, B)$ as (S_1, S_2).
Function $h(A, B) = \sqrt{|A| * |B|} * f$ and f is a cut value of (A, B).
A cut that divides S more evenly with a smaller cut value (i.e. with fewer crossing edges) is chosen.

Figure 7: Document vector division procedure.

Acknowledgements

We thank Herb Chong, James Cooper, Alan Marwick, John Prager, Dragomir Radev, Edward So, and Lillian Lee for helpful discussions, and the anonymous reviewers for their comments and suggestions. Portions of this work were done while the first author was visiting IBM. The first author was partly supported by a McMullen fellowship from Cornell University.

References

Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. 1993. Network flows.

Rie Kubota Ando. 2000. Latent semantic space: Iterative scaling improves inter-document similarity measurement. (To appear).

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the ACL*.

Branimir Boguraev and Mary Neff. 2000. Discourse segmentation in aid of document summarization. In *Proceedings of Hawaii International Conference on System Sciences (*HICSS-33*), Minitrack on Digital Documents Understanding*, Maui, Hawaii. IEEE.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science*, 41:391–407.

Matthias Hemmje, Clemens Kunkel, and Alexander Willett. 1994. LyberWorld – a visualization user interface supporting fulltext retrieval. In *Proceedings of SIGIR'94*, pages 249–260.

Eric P. Jiang and Michael W. Berry. 1998. Information filtering using the Riemannian SVD (R-SVD). In *Proceedings of 5th International Symposium, IRREGULAR'98, Solving Irregularly Structured Problems in Parallel*, pages 386–395.

Teuvo Kohonen. 1997. Exploration of large document collections by self-organizing maps. In *Proceedings of SCAI'97*, pages 5–7.

Robert R. Korfhage and Kai A. Olsen. 1995. Image organization using vibe a visual information browsing environment. In *Proceedings of SPIE*, volume 2606, pages 380–388.

Krista Lagus, Timo Honkela, Samuel Kaski, and Teuvo Kohonen. 1996. Self-organizing maps of document collections: A new approach to interactive exploration. In *Proceedings of Second International Conference on Knowledge Discovery & Data Mining*, pages 238–243.

Xia Lin. 1993. Map displays for information retrieval. *Information Processing & Management*, 29(1):69–81.

Inderjeet Mani and Eric Bloedorn. 1999. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):1–23.

Kai A. Olsen, Robert R. Korfhage, Kenneth M. Sochats, Michael B. Spring, and James G. Williams. 1993. Visualization of a document collection: The VIBE System. *Information Processing & Management*, 29(1):69–81.

Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Earl Rennison. 1994. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. *presented at the ACM Symposium on User Interface Software and Technology, Marina del Rey, CA, November*, pages 2–4.

Gerald Salton and M.J. McGill. 1983. Introduction to modern information retrieval.

J. A. Wise, J.J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. 1995. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proceedings of Information Visualization*, pages 51–58.