디지털도서관 구축과정에서 TREC 텍스트 문서의 시각적 표현에 관한 연구

A Study on the Visual Representation of TREC Text Documents in the Construction of Digital Library

정 기 태(Ki-Tai Jeong)*
박 일 종(Il-Jong K. Park)**

초 록

이용자들은 유사문서를 검색할 때, 각 가지 문서의 시각적표현을 통하여 도움을 얻게 되며 모든 정보검색에 관한 연구는 이용자들의 다양한 요구를 충족시키기 위한 여러 가지의 해결책을 제시하고 있다. 제안되어진 해결책은 알파벳 순서로 만들어 진 파피루스 문서로부터 카드목록, 마이크로 필름을 이용한 저장, 컴퓨터 디스크를 이용한 파일 보관 등에 이르기까지 다양한 방법들을 들 수 있을 것이다. 또한 대부분의 정보검색 시스템들은 Document Surrogate(문헌을 대체할 수 있는 것들), 즉 요약문, 목차, 초록, 리뷰한 내용, 기계가독형목록(MARC) 기록물 등과 같은 서지자료들을 전체논문을 대체하여 이용하게 된다.

본 논문에서는 또 다른 형태의 Document Surrogate로서 용어 리스트의 집단화 방법을 이용해서 찾아보았다. 이 Document Surrogate들은 Multidimensional Scaling (MDS)을 이용해서 2차원 그래프 위에 좌표로써 표현되어지고 있다. 사용된2 차원의그래프 위에서 좌표간의 거리는 문헌들의 유사성을 나타낸다고 해석할 수 있으며 거리가 가까우면 가까울수록 두 문서는 더욱 유사한내용을 포함하고 있다고 해석할 수 있는 것으로 밝혀졌다.

ABSTRACT

Visualization of documents will help users when they do search similar documents, and all research in information retrieval addresses itself to the problem of a user with an information need facing a data source containing an acceptable solution to that need. In various contexts, adequate solutions to this problem have included alphabetized cubbyholes housing papyrus rolls, microfilm registers, card catalogs and inverted files coded onto discs. Many information retrieval systems rely on the use of a document surrogate. Though they might be surprise to discover it, nearly every information seeker uses an array of document surrogates. Summaries, tables of contents, abstracts, reviews, and MARC recordsthese are all document surrogates. That is, they stand infor a document allowing a user to make some decision regarding it, whether to retrieve a book from the stacks, whether to read an entire article, etc.

In this paper another type of document surrogate is investigated using a grouping method of term list. Using Multidimensional Scaling Method (MDS) those surrogates are visualized on two-dimensional graph. The distances between dots on the two-dimensional graph can be represented as the similarity of the documents. More close the distance, more similar the documents.

Keywords: 문헌 대체, 문헌간 유사성, 정보검색시스템, DL, MDS, 용어집단화방법
Document Surrogate, Document Similarity, IRS, DL,
Multidimensional Scaling (MDS), Term Grouping Method

---

* Assistant Professor University of Oklahoma School of Library and Information Studies(ktjeong@ou.edu)
** 계명대학교 문헌정보학과 부교수(ipark@kmu.ac.kr)
■ 논문접수일자 : 2004년 6월 22일
■ 게재확정일자 : 2004년 9월 16일

1. Introduction

All research in information retrieval addresses itself to the problem of a user with an information need facing a data source containing an acceptable solution to that need. In various contexts, adequate solutions to this problem have included alphabetized cubbyholes housing papyrus rolls, microfilm registers, card catalogs and inverted files coded onto discs. The present situation, dominated as it is by the promises and limitations of processors, RAM and hard-disc space, finds many researchers interested in the particular problems posed by large sets of documents existing electronically as text files. Since users cannot physically search or browse electronic text files, they must rely on a machine search.

This paper will report the results of a project designed to measure the effectiveness of a document representation consisting of termgroup-documents called 'Word Code' in reproducing the relationships within a document set as revealed using the technique of multidimensional scaling. As it was mentioned, Word Code as another type of surrogate represents document, and it will contribute in saving memory and processor time in retrieving of digital library. In addition, Word Code will be used as features to represent text documents and will be combined with image features extracted from image. Combined features will represent multimedia documents using image and text. In the multimedia retrieval system, those features will be used to retrieve similar multimedia documents.

A finding that termgroup-document surrogates effectively reproduce the relationships found in the simple term representations would mean a significant savings in memory and processor time. Since searching full text is ponderous at best, some kind of representation of a document might be created. If the document representation could be shown to be similar to a satisfactory degree to the document itself, and if the document representation was more easily searched by the machine, great headway would have been made toward retrieving desired information from the system.

In the context of physical information packages such as books, such document representations, or surrogates, are ubiquitous. The massive volume of information represented by the books in a large research library would itself act as a fatal inhibitor to any attempt to retrieve a particular subset of the information it contained. Fatal, that is, but for the fact that surrogates for the information contained in each book exist in the form of highly formalized MARC records in the library catalog.

Since words are the ingredients in text documents, one of the most basic surrogates would be a list of the words used in a document. In a collection of documents, such a list would comprise a matrix where rows would represent documents and columns would represent the terms they contain (See Table 1). An obvious disadvantage to this scheme becomes evident when one considers the size of the matrix implied by a collection equivalent to, say, one million books averaging (extremely conservatively) one thousand unique words. 1,000,000 X 1,000 = a matrix composed of 1,000,000,000 cells, a considerable burden for most systems (Khorfage, 1997). However, if 1000 terms can be grouped into 10 groups, then the matrix would be reduced to 10,000,000 cells.

〈Table 1〉 Example of a term-document matrix.

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8	Term9	Term10
Doc1	X		X		X	X	X	X		
Doc2										
Doc3	X		X	X		X		X		
Doc4									X	X
Doc5			X				X			
Doc6	X			X	X			X		
Doc7										X
Doc8								X		
Doc9										X
Doc10	X		X				X			

2. Review of the Literature

Begun in 1992, the Text Retrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA), is an ongoing effort to encourage and promote research into the information retrieval problems associated with large full-text document collections (See http://trec.nist.gov/). It represents the first large-scale effort to conduct experiments on such databases. TREC makes available to researchers document sets and related queries; the document sets have been manually examined for relevance to the queries, providing an invaluable control group used to compare retrieval systems. In fact, the ability of researchers to access a controlled and well-defined set of text documents is the primary advantage of adopting TREC document sets as the test set for information retrieval research. In a

discussion of the first two TREC conferences, Khorfage (1997) notes that most participating groups returned similar results, as evaluated in terms of precision and recall. The similar results were obtained in spite of differences in whether the queries were generated manually or automatically, whether negation was employed in the queries, and if so whether the negated term was excluded from the search, included in it, or ignored. Lancaster (1998) takes these findings to suggest that the problems associated with information retrieval in full-text situations were not as severe as some had imagined.

The journal Information Processing and Management devoted an entire issue to a discussion of the sixth TREC conference (TREC-6). In that issue, Voorhees and Harman (2000) give a good introduction to the various tasks and tracks associated with those institutions that fully participate, and discuss the general findings. In addition to offering a digest of the participants' performance on each of the TREC tasks, they address the overall performance of retrieval systems, reporting that since the early iterations of the conference, the effectiveness of the retrieval systems has roughly doubled. Jones (2000) goes beyond reporting TREC-6 results to offer her reflections on the enterprise as a whole. She gives attention to the effectiveness measures traditionally used to evaluate IR systems, Precision and Recall. Precision refers to the degree to which the results retrieved in response to a query are relevant to the information need, and Recall refers to the degree to which all the relevant documents in the collection have been retrieved. These measures are typically seen as expressing an inverse relationship, that is, the greater the precision the poorer the recall.

As useful as these measures are for evaluating the effectiveness of a retrieval system in returning usable information in response to a query, they are not appropriate measures to apply to the current project where the aim is to create and evaluate various document surrogates.

Multidimensional scaling (MDS) is a technique to create a visual display of the relationships in data sets. Young and Hamer (1987) offer this explanation:

> The term multidimensional scaling refers to a family of data analysis methods, all of which portray the data's structure in a spatial fashion easily assimilated by the human eye. That is, they construct a geometric representation of the data, usually in a Euclidian space of fairly low dimensionality. Some multidimensional scaling methods display the data structure in non-Euclidian spaces, and some methods provide additional information about

> how the data structure varies over time, individuals, or experimental conditions. The essential ingredient defining all multidimensional scaling methods is the special representation of data structures.

As might be expected by the ecumenical nature of these remarks, MDS has found applications in a variety of fields. Gazda and Mobely(1994) offer a lengthy discussion of its applications, advantages and disadvantages in conducting sociometry. Bartolucci(1986) notes that MDS has appeared as a useful tool in such varied fields as political and behavior science and archeology.

Rorvig and Fitzpatrick (1997) has been interested in investigating applications of MDS in the field of Information Retrieval. In his article, he submits a TREC document set to MDS in an effort to gain an understanding of the relationships and degree of similarity between the documents. He offers a lengthy discussion of the usefulness of MDS in the TREC context, emphasizing that it is unique in affording researchers the ability to take in the similarity of a document set at a glance. Rorvig, Sullivan, and Oyarce (1998) frame a question that has direct bearing on this research:

> The question thus arises whether a feature vector would be able to recover the initial visual shape created from full-text. A robust feature vector should enable recovery of this shape if it is to be of value in further manipulation of the visual field.

In the context of that remark, his "feature vectors" were representations of the documents using a stemmer to extract word roots; in this project the "feature vectors" will consist of termgroup-documents. Rorvig and Fitzpatrick (2000) discuss MDS in Information Retrieval as well. There they used MDS to construct a control group, i.e., the document set in full text. They were then able to create various treatment groups consisting of the document set after analysis into text categories of different sizes, after application of a stemmer, etc. The procedure in these papers lays the foundation for MDS as a valuable tool in evaluating the effectiveness of feature vectors as document surrogates.

3. Methodology

3. 1 Term Extraction

Since 1990, TREC data has been widely used in a research to develop and evaluate text retrieval systems. Terms extracted from the test set of TREC data comprise a super term list. As a test set of TREC data, 100 documents among 10,000 docu-

ments were randomly chosen as a test set and 6,287 terms were extracted from the given test set. Table 2 is an example of a text document and Table 3 is the Perl program developed by Oyarce in 1999 used to extract terms from the test set. Table 4 is a part of super term list from among the 6,287 terms extracted from the test set. Also, the procedure to extract words from the test data set, to derive Word Code from the document-word matrix, and to draw MDS is given by the flowchart, Figure 1.

〈Table 2〉 An Example of a TREC Document (AP890109-0313)

<DOCNO> AP890109-0313 </DOCNO>
<FILEID>AP-NR-01-09-89 1035EST</FILEID>
<FIRST>u f PM-Britain-GEC 01-09 0556</FIRST>
<SECOND>PM-Britain-GEC,0578</SECOND>
<HEAD>Government Looks at Possible Bid for British Electronics Giant</HEAD>
<DATELINE>LONDON (AP) </DATELINE>
<TEXT>
The government said today that it was looking at a possible bid for the electronics giant General Electric Co. PLC that an international consortium is expected to launch within days. The takeover, which analysts say could be worth between $11.5 billion and $14.2 billion, would be the largest in Britain. The consortium is expected to include Plessey PLC, another electronics company which is the target of a $3 billion hostile takeover bid from GEC and Siemens AG of West Germany, another electronics company. Although no bid for GEC has been formally launched, the Office of Fair Trading has legal powers to look at a bid "in contemplation." "We really are looking at the situation to see who the participants are involved before we can take real active steps," said a </TEXT>
</DOC>

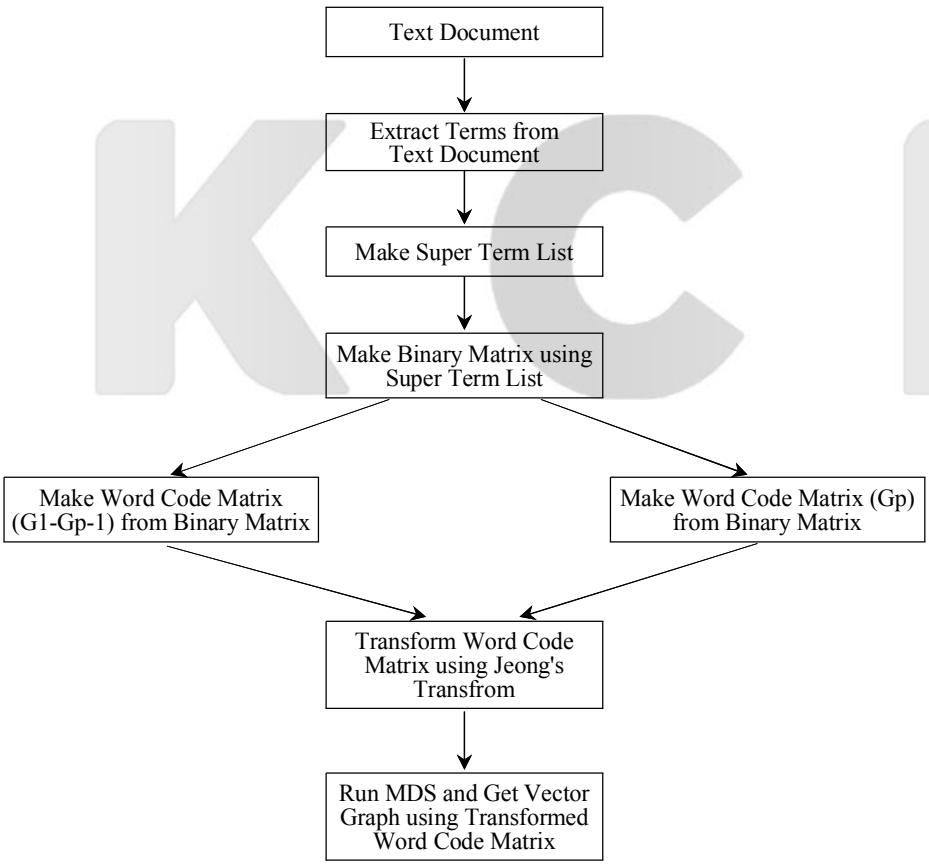
〈Table 3〉 Term Extraction Program using Perl Language

```
##################   MAIN  ##################
&getWORKfiles;          #  Reads filenames in input stream
&getSTOPw;              #  Reads Stop Words
foreach(@INfileList)    #  Starts proc. on input stream
  {
   $SEEin=$_;
   open (IN,"<".$SEEin);              # open file is input stream
   while (<IN>)              # get in documents to id TEXT field
     {
       $t=$_;
            &UPDATEgd;
            &LOCALcount;
     }
close (IN);
  }
&UPDATEgd;
&printGD;
exit;

## Subroutines are excluded ##
```

〈Table 4〉 A Part of Super Term List with Frequencies of Term

```
threat:1
depression:1
regarding:1
wildlife:1
razed:1
evening:1
maintained:1
depositary:1
passenger:1
maximize:1
substantially:1
estimate:1
freeman:1
running:1
samuel:1
leave:1
independently:1
radio:1
paying:1
```



〈Figure 1〉 Flowchart for Text Document Process

3. 2 Binary Matrix

The extracted terms comprising the super term list can be represented in a number of ways, for example by using term frequencies or binary representation. The binary representation method was chosen for this paper because of its simplicity. A super term list was made from the entire document set using the term extraction method. Using a Perl program (See Table 6) to determine whether a particular term existed in a given document, a term-document binary matrix was produced which contained a combination of 0's and 1's. In the binary matrix, a "0" means that the term does not appear in the document and a "1" means that the term does appear within the document. In Table 5, A_1, A_2, …, An represent terms extracted from the given TREC text document set and D_1, D_2, …, Dm represent document from the given TREC text document set. For this experiment, 100 documents from TREC data set were chosen randomly and 6287 terms were extracted from the 100 documents. Table 6 shows a Perl program to generate a binary matrix like Table 5 using a super term list.

〈Table 5〉 An Example of a Binary Matrix from TREC Text Documents

Term / Doc	A_1	A_2	..	..	..	..	..	..	..	A_n
D_1	1	0	..	..	..	..	..	..	..	0
D_2	0	0	..	..	..	..	..	..	..	0
..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..
D_m	0	0	..	..	..	..	..	..	..	0

〈Table 6〉 A Binary Matrix Generating Program using Perl Language

```
###################      MAIN     ###################
open(MatOUT,">$ofil");
&getWORKfiles;          #  Reads filenames in input stream
&getFeatureW;           #  Reads Feature Words
&clearMat;              #  Clear Matrix
foreach(@INfileList)    #  Starts proc. on input stream
  {
   $DocNameRead=$_;
   $SEEin=$_;
   open (IN,"<".$SEEin);        # open file is input stream
   &clearMat;
   while (<IN>)                 # get in documents to id TEXT field
     {
          &makeMat;
     }
   &printMat;
   close (IN);
  }
close(MatOUT);
exit;

## Subroutines are excluded ##
```

3.3 Word Code

For the first experiment, the 6287 terms were divided into Word Codes (i.e. groups of terms) of 10 terms with each group having the frequencies of "1" in that group (See Table 7). For the second experiment, the 6287 terms were divided into Word Codes (WC) of 20 terms (See Table 8) with each group having the frequencies of "1" in that group. Groups of 50 WC (See Table 9) and groups of 100 WC (See Table 10) were also made using the method. In the tables, Gp means total number of words in each document. Word Code is meaning a word except that is a group of words. On that reason, small number of WCs is better than large number of WCs. For example, if 10 WCs can represent a document, it is better to use 100 WCs to represent same document. It is a lot better than using 6287 words to represent same document without using WCs.

〈Table 7〉 An Example of 10 Terms Word Code Matrix from the Binary Matrix

WCM / Doc	G_1	G_2	..	..	..	..	..	..	..	G_P
D_1	1	0	..	..	..	..	..	..	..	213
D_2	1	0	..	..	..	..	..	..	..	276
..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..
D_m	3	0	..	..	..	..	..	..	..	256

〈Table 8〉 An Example of 20 Terms Word Code Matrix from the Binary Matrix

WCM / Doc	G_1	G_2	..	..	..	..	..	..	..	G_P
D_1	1	1	..	..	..	..	..	..	..	213
D_2	1	0	..	..	..	..	..	..	..	276
..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..
D_m	3	0	..	..	..	..	..	..	..	256

〈Table 9〉 An Example of 50 Terms Word Code Matrix from the Binary Matrix

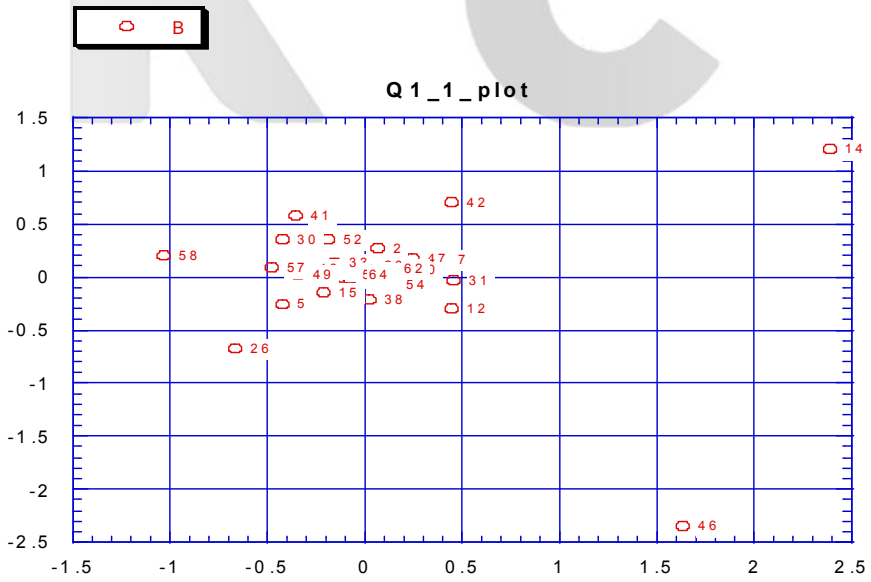
WCM / Doc	G_1	G_2	..	..	..	..	..	..	..	G_P
D_1	2	2	..	..	..	..	..	..	..	213
D_2	2	3	..	..	..	..	..	..	..	276
..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..
D_m	3	5	..	..	..	..	..	..	..	256

〈Table 10〉 An Example of 100 Terms Word Code Matrix from the Binary Matrix

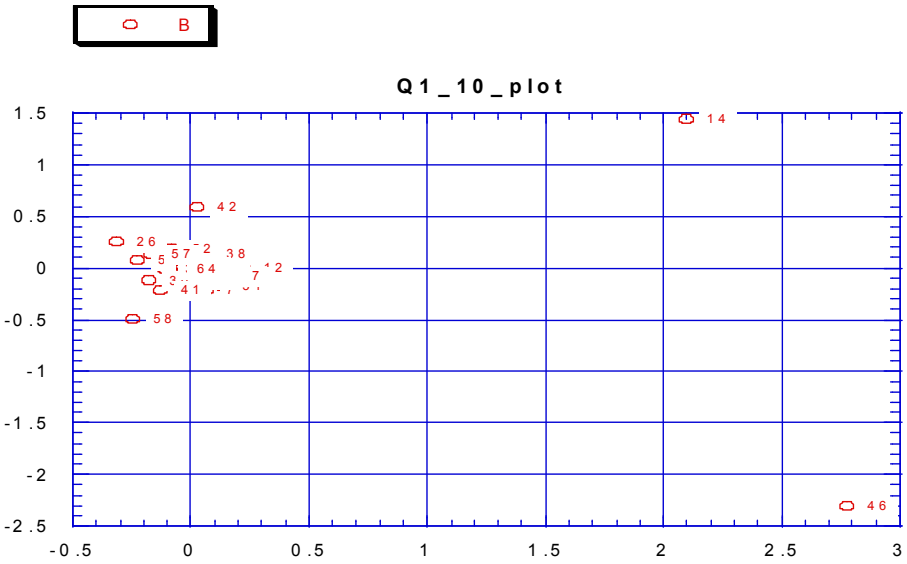
Doc \ WCM	G_1	G_2	..	..	..	..	..	..	..	G_P
D_1	4	3	..	..	..	..	..	..	..	213
D_2	5	3	..	..	..	..	..	..	..	276
..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..
D_m	8	5	..	..	..	..	..	..	..	256

3. 4 Multi-Dimensional Scale

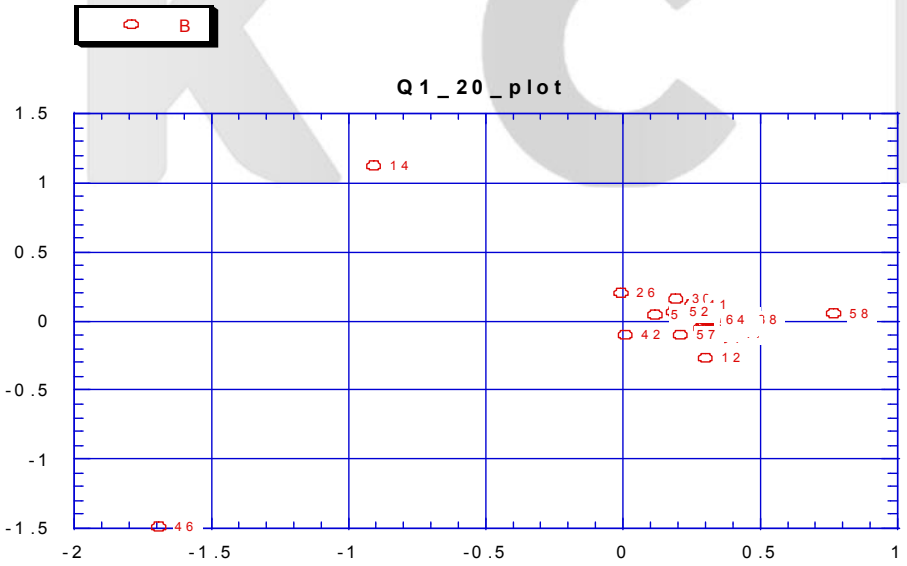
The original binary matrix was used without grouping to create a vector graph using Multi-Dimensional Scaling (MDS) and its vector graph is shown in Figure 2. Next, the 10 term WC, 630 groups in total for this experiment, were used to draw a vector graph using MDS and its vector graph is shown in Figure 3. The same was done for 20 term WC (316 groups in total) and its vector graph is shown in Figure 4, 50 term WC (127 groups in total) and its vector graph is shown in Figure 5, and 100 term WC (64 groups in total) were also used to draw vector graphs using MDS and its vector graph is shown in Figure 6.



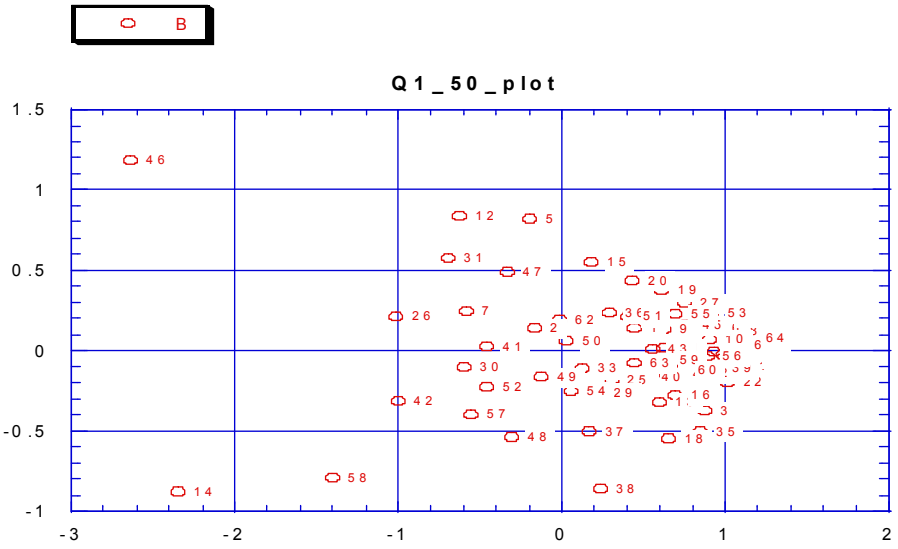
〈Figure 2〉 Vector Graph of Binary Matrix



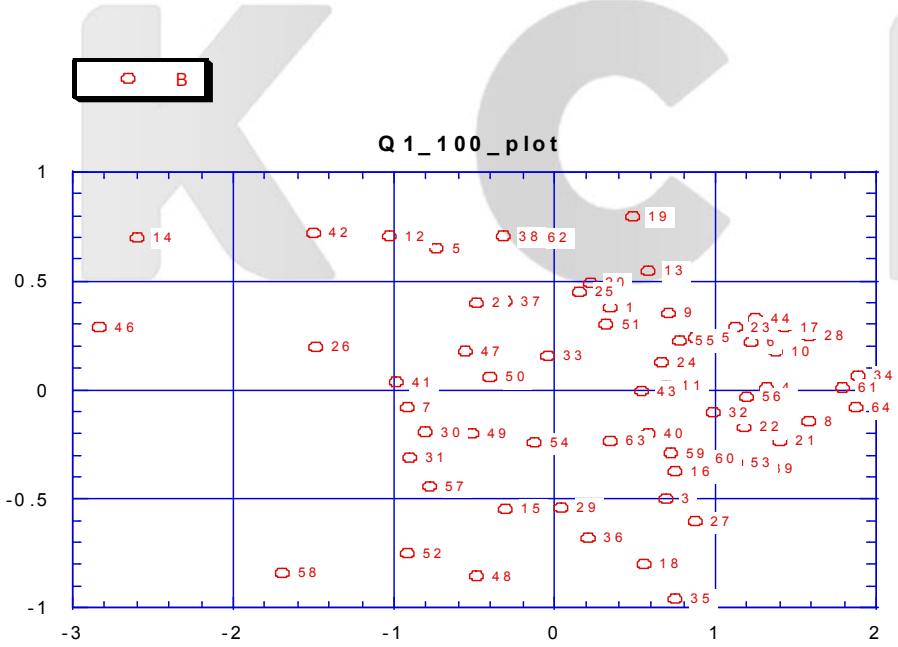
〈Figure 3〉 Vector Graph of 10 Term-Grouping of Word Code



〈Figure 4〉 Vector Graph of 20 Term-Grouping of Word Code



<Figure 5> Vector Graph of 50 Term-Grouping of Word Code



<Figure 6> Vector Graph of 100 Term-Grouping of Word Code

4. Discussion and Conclusion

Visualization in information retrieval (IR) is attractive since it allows for users to see what the IR is doing in a way that is simple to understand. In Figures 2 thru 6 the distances between vectors that are representing documents reveals the similarities between documents. In analyzing the vectors a clustering method is used. If the vectors are clustered, the documents represented by the clustered vectors are very closely related. Seeing the graphs it is clearly noticed that there are two outsiders: document numbers 14 and 46. Even though 100 test documents come from a homogeneous document collection, there are still some documents whose contents are not similar to others.

When we see the shapes of the vector graphs we notice that the Figures 2, 3, and 4 are very similar, but Figures 5 and 6 are somewhat dissimilar at first glance. However, all of them are very similar if you turn around the figures. Figures 5 and 6 depict the similarity of the documents as determined by a Word Code of 50 and 100, respectively. One would expect that as increasingly large Word Codes are used as document surrogates the accuracy with which they represent the documents should begin to suffer. It should have researched more closely in the future.

In two ways this research can contribute in digital library. First, the visualization of documents can be a great help for users in searching electronic text documents. The clusters can be a starting point in retrieving text documents. Especially Word Code can save memory and processing time. If we use Word Code of 100 words, the system may need 100 times less memory than not using Word Code. Second, Word Code will be a key fact in combining with image features for multimedia retrieval system. Currently not many features could be extracted from image using Content Based Image Retrieval technique. To have weight balance between text documents and image documents for multimedia documents retrieval system, extracted words from text documents should be manipulated. Upon this idea, a research will be done in the near future.

The results of this project point toward two main areas for further research. First, a numerical analysis of the similarities between the documents should be undertaken. Such a project would complement the MDS depiction used here. Second, the test set may be expanded to include heterogeneous document. Creating MDS depictions of Word Codes extracted from heterogeneous documents would provide a platform to assess the robustness of the Word Codes in acting as document surrogates.

REFERENCES

박일종. 2000. 디지털 도서관시대에 대비한 도서관자동화시스템의 비교효용성과 개발방향에 대한 연구. 『정보관리학회지』, 17(2); 207-231.

Bartolucci, Alfred. 1986. "Multidimensional Scaling and the Information it Conveys." *American Journal of Public Health,* 76(7); 747-71.

Gazda, George., Mobely, Jerry. 1994. "Multidimensional Scaling for the 21st Century." *Journal of Group Psychotherapy, Psychodrama & Sociometry*. 47(2).

Goodrum, A., Rorvig, M., Jeong, K., Suresh, C. 2000. An Open Source Agenda for Research Linking Text and Image Content Features. *Journal of the American Society for Information Science.*

Jones, Karen Sparck. 2000. "Further re - flections on TREC." *Information Processing and Management,* 36: 37-85.

Khorfage, Robert B. 1997. *Information Storage and Retrieval*. New York: John Wiley.

Lancaster, F.W. 1998. Indexing and Abstracting in Theory and Practice. Champaign, IL: University of Illinois.

Rorvig, M., Fitzpatrick, S. 1997. Visualization and Scaling of TREC topic document sets. International Journal of Information Processing and Management.

Rorvig, M., Sullivan, T., Oyarce, G. 1998. A visualization case study of fea-ture vectorand stemmer effects on TREC topic-document Proceedings of the 1998 Annual Meeting of the American Society for Information Science.

Rorvig, M., Fitzpatrick, S. 2000. "Shape recovery: a visual method for eval-uation of information retrieval experiments." *Journal of the American Society for Information Science*, 51(13); 1205-1210.

Voorhees, Ellen M., Harman, Donna. 2000. "Overview of the Sixth Text Retrieval Conference (TREC-6)." *Information Processing and Management* , 36: 3-35.

Young, F.W., Hamer, R.M. (Eds.) 1987. Multidimensional scaling: History, theory and applications. Hillsdale, NJ: Erlbaum.