Mariusz Grabowski
Paweł Lula

Automatic Identification of Time and Subject Related Patterns in Large Collection of Scientific Publications

Article Info:

Management Information Systems,
Vol. 6 (2011), No. 3,
pp. 003-007

Received 12 November 2010
Accepted 25 June 2011

UDC 659.25:004

Summary

In the past decades, we have witnessed a dynamic development of scientific literature. The most important changes of this phenomenon relate to: (1) continuous increase in the number of published papers; (2) rapid development of electronic means of publication; (3) increased access to publications.

The above trends are to be assessed positively. We must not, however, forget that the ongoing changes bring about certain difficulties that are new to us. Some of these difficulties may concern our inability to get acquainted with the most up-to-date issues because it is impossible to analyze the flood of information represented in numerous publications. The way to solve this problem may be application of automatic tools for information retrieval from text documents and application of data analysis methods that would enable a precise classification of documents based on the facts contained in them.

Several methods connected with time and subject-related classification of text documentation will be presented as well as an attempt of their evaluation. As a result, key subjects as well as the trends that shape the analyzed research discipline will be identified. Papers published in MIS Quarterly in the period starting from 1977 until present will be used as the empirical data set.

Keywords

information retrieval, information systems, text mining, data mining

1. Introduction

Nowadays, the dynamic development of scientific literature can be observed. It manifests itself by:

- Continuous increase in the number of published papers;
- Rapid development of electronic means of publication;
- Increased access to publications.

Apart from all positive aspects, the development in the field of scientific literature also has some negative results. The flood of information, forecasted in Ackoff (1967), is one of them. It seems that applying computed-based tools, especially to filter and compress information, may be helpful in solving that problem. The ideas about using computer tools in the analysis of text documents have a long history; they include a number of approaches, e.g. Chomsky's works on grammar, Turing test, natural language processing, *Eliza* program by Weizenbaum, machine translation and text mining. The text mining (Hearst, 1999) approach is relatively new. It is the use of data mining techniques (Witten & Frank, 2000) on text documents. One of the first definitions of text mining was proposed in Hearst, (2003):

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources.

Text mining methods and tools have been used in this paper to identify key subjects in scientific publications in the field of information systems (IS) in the period from 1977 to 2006, and to analyze the main changes in this period which could be observed in the IS field.

2. The Purpose of Analysis

IS is an interdisciplinary field which is focused, as it was proposed by Lyytinen & King, (2004, p. 221), on:

> *a market of ideas in which scholars (and practitioners) exchange their views regarding the design and management of information and associated technologies in organized human enterprise.*

The interdisciplinary character of information systems is expressed in taking two perspectives into consideration (Laudon & Laudon, 2002, pp. 14-15): (1) technical, which consists of management science, computer science and operational research and (2) behavioral, which incorporates psychology, economics and sociology.

By using text mining approach, the authors of this paper want to verify, if the interdisciplinary character of IS field is reflected in the variety of topics covered. The second research question is connected with the third postulate raised by Lyytinen & King (2004, p. 226) when they discussed the scientific legitimacy of the field. This was the postulate of "the plasticity of the field with respect to changing circumstances." The answer to this question may be found in the analysis of the changes in topics covered over time, presented in this paper.

According to interdisciplinary character of IS, we may not include periodicals that narrow the scope of interest only to certain area, i.e. technical aspects of information technology. For this reason, *MIS Quarterly* (http://www.misq.org) was used as the source of papers, as it is a widely recognized leading scientific magazine of IS field. It fully reflects the whole scope of IS interests. In the opinion of the authors of this article, limiting the research sample to this one title would not disqualify the results of the analysis, as *MIS Quarterly* represents high scientific quality. In the future, however, this analysis may be extended to other periodicals of IS field such as: *Information Systems Research*, *European Journal of Information Systems* and *Information Systems Management* to mention some.

The set of documents which was analyzed during the research included all the papers published in the period from 1977 to 2006. The total number of analyzed papers was 768. In fact, this includes all annuals of *MIS Quarterly* published until 2006. This also allowed us to conduct a time related analysis.

Each paper was described by *title*, *keywords* and *abstract* section. The following tools were used during the research: *STATISTICA* Text Miner, Microsoft Excel spreadsheet and other individually developed computer programs written in Perl and Visual Basic.

3. Document Preprocessing

The preprocessing of documents was the first step in the process of analysis. It included:

- transformation to plain text format;
- stemming - during that process all words were replaced by their basic forms;
- removing irrelevant words from English stop-list;
- limiting the set to the words which appeared in keywords section;
- identifying phrases;
- limiting the set to the terms which appeared with the frequency from 3% to 50%; and
- getting bag-of-words representation.

As the result of the preprocessing we obtained the bag-of-word, which is the most popular form of text documents representation, easy for further processing with data analysis methods. It is a matrix, in which rows correspond with the words (or phrases) appearing in a document, and columns correspond with the documents in the set. Element x_{ij} indicates how many times $i-th$ word is present in $j-th$ document. Basing on bag-of-words we can obtain frequency matrix, which is the starting point for further analysis. This method of document representation is known as vector space model and was first proposed by Salton, Wong, & Yang (1975).

4. Key Subjects Identification

In order to identify key subjects, the following procedure was used:

- Transformation of frequency matrix (vector space model representation of a document) to an inverse document frequency (matrix). During this step, all nonzero values of frequency matrix were transformed according to the formula proposed in Manning & Schütze (1999):

  $$x_{idf}(i,j) = (1 + \log(x_f(i,j))) * \log(N/di)$$

  Where: – total number of documents, – number of documents containing word, – number of occurrences of word in document.

  The purpose of making on the value is merely to draw attention to the fact of the word occurrence and simultaneously to weaken (but not exclude) the importance of information about the number of occurrences of a given word (function "flattens" these values to a great degree). The expression gives more significance to these words that occur in a relatively small number of documents. During the transformation described in this step, values "0" remain unchanged.
- Singular Value Decomposition (SVD) of matrix. SVD is a linear algebra tool, which is a basis to LSI - Latent Semantic Indexing, described also as LSA - Latent Semantic Analysis (Deerwester, Dumais, Fumas, Landauer, & Harsman, (1990)). The main purpose of LSA is to determine a new space,

with a lower dimension in which it would be possible to analyze both: the set of documents and the set of words. SVD (decomposition according to singular values) allows expressing any matrix as the product of three other matrices: The above matrices have the following characteristics:

- ✓ Matrix $\mathbf{S}$ is a diagonal matrix; it can be proved that its elements are square roots of the eigenvalues of matrix $\mathbf{X}_{idf}\mathbf{X}_{idf}^T$ and matrix $\mathbf{X}_{idf}^T\mathbf{X}_{idf}$; these values specify the importance of consecutive dimensions in new space;
- ✓ Elements of matrix $\mathbf{S}$ are in decreasing order;
- ✓ There is a dependence of $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, which means that columns of matrix $\mathbf{U}$ are orthogonal;
- ✓ Columns of matrix $\mathbf{U}$ are eigenvectors of matrix $\mathbf{X}_{idf}\mathbf{X}_{idf}^T$, which means that they determine principal components for the set of words;
- ✓ Values determined as $\mathbf{US}$ are the coordinates of words in a new space;
- ✓ There is a dependence of $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, which means that the columns of matrix $\mathbf{V}$ are orthogonal;
- ✓ Columns of matrix $\mathbf{V}$ are eigenvectors of matrix $\mathbf{X}_{idf}^T\mathbf{X}_{idf}$, which means that they determine the principal components for the set of documents;
- ✓ Values determined as $\mathbf{VS}$ are the coordinates of documents in the new space.

In order to determine the importance of words, the Euclidean distances between the new coordinate origin and the points representing corresponding words were assigned. Those values were obtained through getting the square roots from matrix IMP calculated according to the formula:

$$\mathbf{IMP} = \mathbf{US}(\mathbf{US})^T = \mathbf{USSU}^T$$

The above presented approach was used for processing the collection of abstracts of scientific publications in the field of information systems. As it was mentioned before, it constitutes a set of 768 files, published in *MIS Quarterly* in the period of 1977-2006.

In order to simplify the interpretation of the results, the obtained measures were scaled in such a way that maximal value would be assigned 100% value. The results for 12 most important terms are presented in Fig. 1.

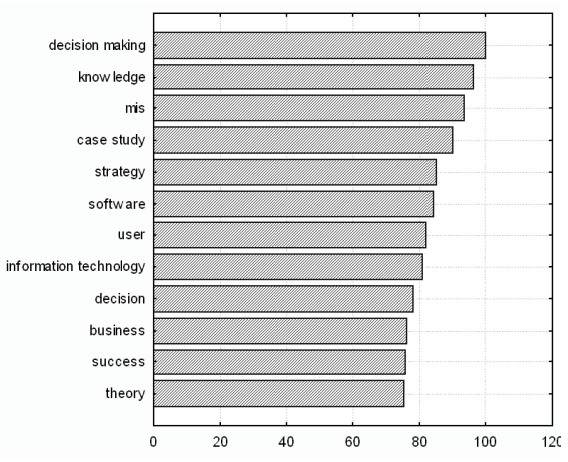

Figure 1 Twelve most frequent used words.

The obtained results confirm the interdisciplinary character of IS field in *MIS Quarterly* publications. Twelve most frequently used terms belong to both perspectives of the discipline - technical and behavioral. The terms *software* and *information technology* belong to the technical perspective, whereas *user*, *success*, *decision* and *decision making* are behavioral in character. The other terms - *knowledge*, *mis*, *case study*, *strategy*, *and business* - lie somewhere in between and connect both perspectives in "organized human enterprise". The high rank of the term *theory* is an evidence of the academic character of IS field. However, the first rank of *decision making* and the last of *theory* show that IS field gives more attention to practice than to theory. Such a set of keywords, representing the topics covered over the years in *MIS Quarterly*, fully complies with the definition of IS field quoted in the beginning of section 1 of this paper.

5. Identification of Changes Over Time

In order to determine the most important changes in the popularity of terms over time, the following calculations were performed:

- transformation of $\mathbf{X}_f$ matrix (frequency matrix) to binary matrix (positive values in frequency matrix were transformed to "1" value; "0" remained unchanged);
- calculation of total number of papers with each key-word in each year of analysis;
- visualization of changes in time for every identified keyword.

Figures 2-4 illustrate the time dependant occurrences of the three selected keywords: "decision making", "information technology" and "knowledge" correspondingly.

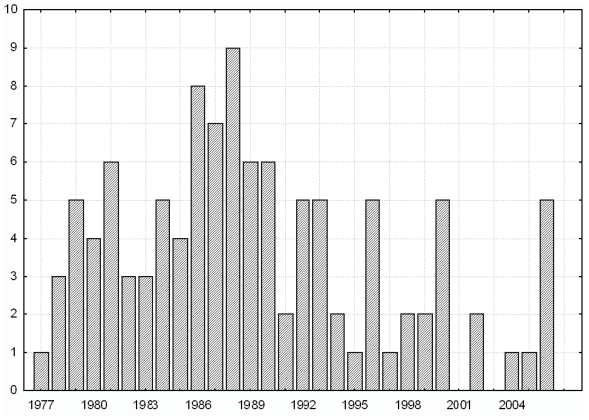

Figure 2 Papers on "decision making" over the analyzed period

We may observe that the keyword "decision making" was the center of attention in the second half of the 1980s and still is the subject of a moderate concern. The keyword "information technology" started to gain significance in the mid-eighties and nowadays is the most popular keyword of the MIS literature.

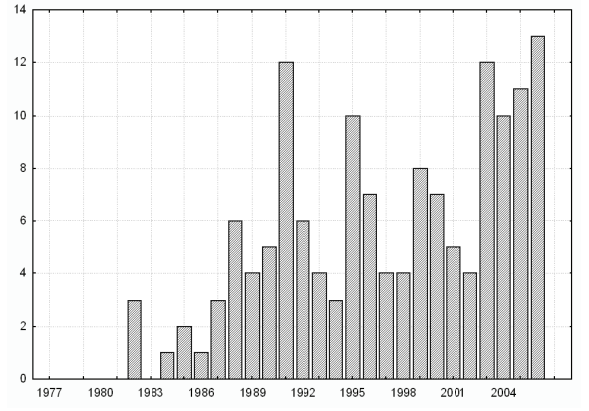

Figure 3 Papers on "information technology" over the analyzed period

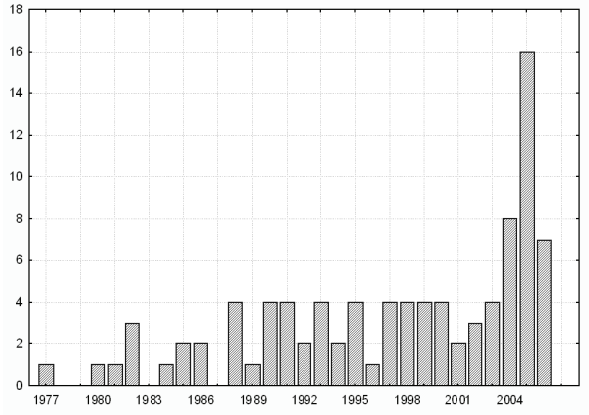

Figure 4 Papers on „knowledge" over the analyzed period

The keyword "knowledge" is rather evenly distributed during the analyzed period; however, in the years 2004, 2005 and 2006 it was in the center of attention.

To simplify the interpretation of the results from table 1, fig. 5 is submitted which summarizes observed irregularities.

Table 1 Results from time-dependent analysis

Year	business	case study	decision	decision making	information technology	knowledge	mis	software	strategy	success	theory	user	Number of papers
1977	1	1	3	1	0	1	8	1	0	0	3	2	15
1978	0	1	2	3	0	0	8	0	3	2	0	4	18
1979	1	1	3	5	0	0	11	3	4	1	0	5	19
1980	1	3	1	4	0	1	5	2	0	2	0	2	19
1981	2	0	4	6	0	1	10	1	5	0	0	2	20
1982	2	2	2	3	3	3	11	2	3	1	1	7	25
1983	4	1	0	3	0	0	6	2	3	3	0	3	20
1984	3	1	2	5	1	1	7	1	0	5	1	6	20
1985	1	1	2	4	2	2	10	0	5	4	0	4	24
1986	1	3	5	8	1	2	7	0	3	3	0	4	26
1987	3	3	5	7	3	0	7	2	6	6	3	3	33
1988	4	3	6	9	6	4	1	6	10	5	0	3	36
1989	7	4	3	6	4	1	7	3	8	2	0	7	29
1990	1	4	1	6	5	4	0	5	7	4	0	1	24
1991	5	4	3	2	12	4	6	5	10	1	2	3	28
1992	6	7	3	5	6	2	7	4	7	6	0	3	25
1993	2	6	4	5	4	4	1	2	3	2	3	5	26
1994	2	6	1	2	3	2	3	5	4	0	1	4	21
1995	8	4	3	1	10	4	4	2	2	2	2	4	25
1996	7	3	4	5	7	1	0	1	2	3	1	2	21
1997	3	5	2	1	4	4	5	1	6	1	2	2	20
1998	2	4	5	2	4	4	2	1	6	4	2	0	21
1999	8	6	2	2	8	4	3	1	3	2	9	4	28
2000	1	9	5	5	7	4	1	7	4	2	9	1	37
2001	1	1	1	0	5	2	2	2	2	2	0	1	31
2002	2	4	1	2	4	3	1	2	5	2	5	0	26
2003	2	3	3	0	12	4	1	1	3	2	4	2	28
2004	5	3	0	1	10	8	1	2	3	1	4	4	29
2005	3	3	2	1	11	16	1	1	4	1	8	2	35
2006	5	5	5	5	13	7	1	8	4	2	9	6	39
Total	93	101	83	109	145	93	137	73	125	71	69	96	

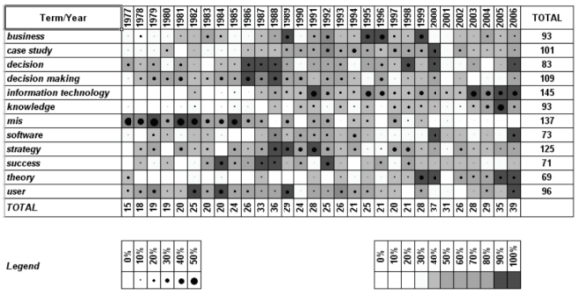

Figure 5 Interpretation of results from time-dependent analysis

Rows correspond to terms and columns correspond to years. Black circles indicate the

percentage of papers published in a corresponding year and containing a given word. The level of gray of the corresponding field defines the level of popularity of a given term in a corresponding year. It is calculated as a fraction of the number of documents published in a given year and a maximal number of the documents containing a given term observed over the period.

The observed changes in the topics illustrate plasticity of is field, which attempts to find theoretical base for the practical decisions made in "organized human enterprise". We may see that, in the initial period of analysis, approx. 40-50% occurrences contain the term "mis". In the recent years, terms connected with *knowledge* and *information technology* have gained popularity.

Such order of terms, correlated with the topics covered over the years, corresponds with the observed trend. Information systems tend to be more and more distributed. This trend shifts the interest of the designers and information system builders from monolithic (such as MIS) to more inter-operational systems, in which the accent is placed on knowledge exchange and high IT integration.

6. Conclusion

The text mining approach applied to the information retrieval from a large collection of scientific papers, presented in this paper, proved its usefulness.

The obtained results seemed to confirm the theses formulated in the beginning of this article. IS Field, as reflected in *MIS Quarterly* publication, proved its interdisciplinary character and plasticity. The conducted analysis also allowed us to formulate the following methodological remarks:

- Algorithms of key-words identification have crucial meaning in IR problems;
- Time related aspects of text documents analysis may be a useful approach to statistical modeling and forecasting.

The presented research can be extended in the future to other collections of publications or/and other aspects of text analysis.

References

Ackoff, R. L. (1967). Management Misinformation Systems. *Managament Science , 14* (4), 147-156.

Deerwester, S., Dumais, S., Fumas, G., Landauer, T., & Harsman, R. (1990). Indexing by latent semantic analysis. *Journal for the American Society for Information Science , 41* (6), 391-407.

Hearst, M. (1999). *Untangling Text Data Mining.* Retrieved March 20, 2007, from School of Information, University of California, Berkeley: http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html

Hearst, M. (2003, October 17). *What is Text Mining?* . Retrieved March 20, 2007, from School of Information, University of California, Berkeley: http://people.ischool.berkeley.edu/~hearst/text-mining.html

Laudon, K. C., & Laudon, J. P. (2002). *Management Information Systems. Managing the Digital Firm.* Upper Saddle River: Prentice-Hall.

Lyytinen, K., & King, J. L. (2004). Nothing At The Center? Academic Legitimacy in the Information Systems Field. *Journal of the Association for Information Systems , 5* (6), 220-246.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press.

Salton, G., Wong, A., & Yang, C. S. (1975)). A Vector Space Model for Automatic Indexing. *Communications of the ACM , 18* (11), 613-620.

Witten, I. H., & Frank, E. (2000). *Data mining.* New York: Morgan-Kaufmann.

Mariusz Grabowski
Cracow University of Economics
Department of Computational Systems
ul. Rakowicka 27
31-510 Kraków
Poland
Email: Mariusz.Grabowski@uek.krakow.pl

Paweł Lula
Cracow University of Economics
Department of Computational Systems
ul. Rakowicka 27
31-510 Kraków
Poland
Email: Pawel.Lula@uek.krakow.pl