An Application of Latent Semantic Analysis for Text Categorization

Article *in* International Journal of Computers, Communications & Control (IJCCC) · April 2015

DOI: 10.15837/ijccc.2015.3.1923

CITATIONS
9

READS
2,058

2 authors:

Gang Kou
Southwestern University of Finance and Economics
265 PUBLICATIONS **6,802** CITATIONS
SEE PROFILE

Yi Peng
Baidu Online Network Technology
28 PUBLICATIONS **315** CITATIONS
SEE PROFILE

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 10(3):357-369, June, 2015.

An Application of Latent Semantic Analysis for Text Categorization

G. Kou, Y. Peng

Gang Kou
School of Business Administration
Southwestern University of Finance and Economics, Chengdu, China
No.555, Liutai Ave, Wenjiang Zone
Chengdu, 611130, China
kougang@swufe.edu.cn

Yi Peng*
School of Management and Economics
University of Electronic Science and Technology of China, Chengdu, China
No.2006, Xiyuan Ave, West Hi-Tech Zone
Chengdu, 611731, China
*Corresponding author: pengyicd@gmail.com

Abstract: It is a challenge task to discover major topics from text, which provide a better understanding of the whole corpus and can be regarded as a text categorization problem. The goal of this paper is to apply latent semantic analysis (LSA) approach to extract common factors that representing concepts hidden in a large group of text. LSA involves three steps: the first step is to set up a term-document matrix; the second step is to transform the term frequencies into a term-document matrix using various weighting schemes; the third step performs singular value decomposition (SVD) on the matrix to reduce the dimensionality. The reduced-order SVD is the best k-dimensional approximation to the original matrix. The experiment uses more than fifteen hundreds research paper abstracts from a specific field. Because different factor solutions of the LSA suggest different levels of aggregation, this work examines thirteen solutions in the experiment. The results show that LSA is able to identify not only principle categories, but also major themes contained in the text.
Keywords: Latent Semantic Analysis, Topic extraction, Text Mining, Information Retrieval.

1 Introduction

Many multidisciplinary fields, such as data mining, bioinformatics, biochemistry, and neuroscience, emerge in the past several decades. Since multidisciplinary fields involve theories, methods, and techniques from multiple disciplines, it is not easy to comprehend all the research efforts in these fields. Text categorization, which organizes documents into groups based on their underlying structures, can help capturing the large amount of activities and diversity of a multidisciplinary field.

The goal of this paper is to apply latent semantic analysis (LSA) approach to detect major research topics and themes of a multidisciplinary field. In particular, it is intended to address three questions: what are the core research areas of the selected field, what are the major research themes, and what is the dynamics of the discipline? LSA is an automatic mathematical and statistical technique for uncovering common factors that representing concepts hidden in text[1,2,3,4]. Previous investigations in psychology and computer science have proved that LSA resembles the way the human brain distills meaning from text and is capable of inferring much deeper relations in the text data[3,5].

The rest of the paper is organized as follows. Section 2 describes the basic concepts of LSA. Section 3 presents the experimental study that was used to identify the core research areas

and themes. Section 4 discusses the results of this analysis, focusing on three important factor solutions of LSA. Section 5 summarizes the paper with conclusions and limitations.

2 Research method

Latent semantic analysis (LSA) is a theory of knowledge acquisition, induction and representation[2]. It was first introduced as an information retrieval (IR) technique by [1] and [6]. It is an automatic mathematical learning technique for analyzing the relationships and similarity structures among documents and terms, relying on no human experiences, prior theoretic models, semantic dictionaries, or knowledge bases[3].

Similar to factor analysis, principal components analysis, and linear neural networks, the main purpose of LSA is dimension reduction, which is realized through a matrix operation called singular value decomposition (SVD). SVD is a means of decomposing a matrix indo a product of three simpler matrices. By retaining the k largest singular values, the resulting reduced-order SVD provides the best k-dimensional approximation to the original matrix, in the least square error sense[7]. In the results of SVD, two sets of factor loadings, one for the words and one for the documents, are generated. Each term and document is represented as a k-dimensional vector in the same latent semantic space derived by the SVD. Thus each latent semantic factor is now associated with a collection of high-loading terms and high-loading documents[5]. High-loading terms and documents are used to interpret and label the corresponding factor. The number of factors is an input parameter that needs to be provided before SVD computation. As the number of factors changes, LSA groups key terms or documents into various levels of aggregation. When it is applied to identify important topics of a certain discipline using a collection of representative papers, a higher level of aggregation (e.g., 2 factors) indicates key research areas and a lower level of aggregation (e.g., 100 factors) represents general research themes[5].

The LSA analysis can be summarized in three main steps. The first step is to set up a term-document matrix in which each row stands for a key word or term and each column stands for a document or context in which the key word appears. An entry in the matrix is the frequency of a key word in the corresponding document. The second step is to transform the term frequencies in a term-document matrix using various weighting schemes. The third step is to perform SVD on the matrix to reduce the dimensionality, which is the key feature of the LSA method. In this step only the k largest singular values are retained. The reduced-order SVD is the best k-dimensional approximation to the original matrix[7].

Extensive experiments have demonstrated that the classification performance of LSA is robust[8] and it is capable of inferring relations in the text [3,5]. It can be used in information retrieval (IR), search optimization, classification, clustering, filtering and other IR-related applications[7]. Readers interested in mathematical details of the LSA approach can refer to [1].

3 Experimental study

This section describes the data source and the implementation details of LSA analysis that is utilized to identify the core research areas and research themes for the selected field.

3.1 Data sources

The field of Multiple Criteria Decision Making (MCDM) and multiattribute utility theory (MAUT) has grown exponentially and made remarkable progress since 1960s. As a multidisciplinary field, MCDM/MAUT has close collaboration with some neighboring disciplines, such as

mathematical programming, organizational behavior, engineering, decision analysis, and negotiation science[9]. During the past twenty years, extensive research papers have been published in MCDM, MAUT, and related disciplines. In the experiment, LSA is applied to a collection of MCDM/MAUT publications to extract major research topics and identify the trends of the field.

Since previous studies, such as [10], [11] and [12], have investigated the major areas and the evolution of MCDM/MAUT before 1990s, articles published before 1985 were not included in the analysis. A total of 1515 research abstracts published in 16 refereed MCDM-related journals in the English language during the period of 1985 to February 2009 that contain key words: multiple criteria and multicriteria, were collected. As the first and unique journal in multiple criteria decision analysis, articles published in the Journal of Multi-Criteria Decision Analysis were all collected (from 1992 through 2007).

The 16 refereed MCDM journals were selected according to two criteria: (1) journals appeared frequently in the Multiple Criteria Decision Aid bibliography on the International society on MCDM website[13]; (2) the most relevant and top-rated MCDM journals listed by [14] and [15]. Each article collected in the dataset is stored in Microsoft Excel as one row with five fields: article title, author(s), journal name, year of publication, and abstract.

Table 1 lists the journals and the number of abstracts included in the text data. About 34% of the articles were published in the *European Journal of Operational Research*, with about 19% in the *Journal of Multi-Criteria Decision Analysis* and 8.5% in the *Journal of the Operational Research Society*.

Table1. Refereed MCDM journal articles, 1985-2008

Journals	Number of Articles
European Journal of Operational Research (EJOR)	519
Journal of Multi-Criteria Decision Analysis (JMCDA)	292
Journal of the Operational Research Society (JORS)	130
Computers & Operations Research (C&OR)	88
Fuzzy Sets and Systems (FSS)	86
Computers & Industrial Engineering (C&IE)	70
Decision Analysis (DA)	64
Omega	64
Mathematical and Computer Modelling (M&CM)	41
Annals of Operations Research (AOR)	40
Decision Support Systems (DSS)	35
Management Science (MS)	31
Operations Research (OR)	18
Journal of Optimization Theory and Applications (JOTA)	16
Theory and Decision (TD)	11
Organizational Behavior and Human Decision Processes (OBHDP)	10

Figure 1 summarizes the number of publications in the field of MCDM from 1985 to February 2009. Because text data were retrieved in October 2008, the number of abstracts collected for the year of 2009 can not reflect the real publication trend and therefore is ignored in Figure 1. As seen in Figure 1, the MCDM publications have been increased rapidly since 1992 and the number of MCDM publications has increased 4.7 times from 1985 to 2008.

3.2 Text preprocessing

The initial step of LSA analysis is to represent the text as a term-document matrix in which each row stands for a term and each column stands for a document. In order to set up such a matrix, this study started the analysis with text preprocessing procedures that are popular in the information retrieval and text mining[16,17].

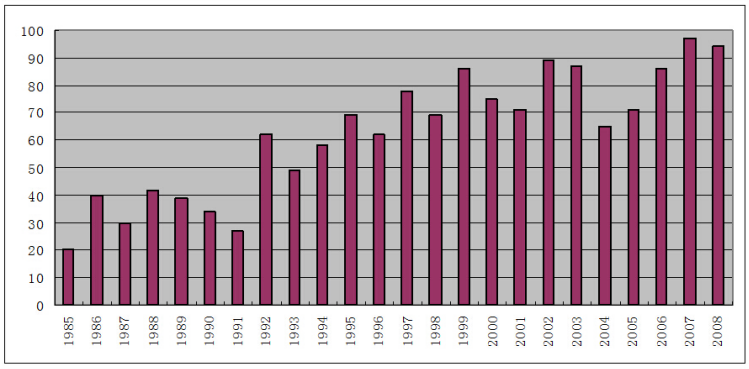

Figure 1: MCDM Publications from 1985 to 2008

The text preprocessing procedure consists of tokenization and term reduction. Tokenization divides documents into a set of terms. In this study, each article is represented by its title and abstract. Since titles are informative of research papers and normally contain pertinent key words, the weight of titles is set twice as much as abstracts. The 1,515 MCDM research papers generated a dictionary of 9,322 terms. Tokenization was implemented using a self-developed C++ program.

3.3 Term frequency matrix

Text preprocessing produced a term-frequency matrix with 1,515 columns (papers) and 3,299 rows (terms). Originally, an entry in the matrix contains the number of times a term occurs in a document. A term-frequency matrix measures the association of a term with respect to a given document[17]. There are many methods to define term weights. In this study, the *tf-idf*, a traditional term-frequency weighting, was used to transform the raw term frequencies in the matrix. The *tf-idf* weighting scheme combines term frequency (TF) and inverse document frequency (IDF) together:

$$W_{ij} = tf_{ij} \times idf_i \tag{1}$$

where tf_{ij} is term frequency and idf_i is the inverse document frequency of term i.

Inverse document frequency represents the importance of a term and is defined as:

$$idf_i = log_2(N/df_i) + 1 \tag{2}$$

N is the total number of documents and df_i is the document frequency of term i.

IDF implies that the discriminative power of a term will be decreased if it occurs in many documents. In other words, the importance of a term will increase if it appears in a limited number of documents. The reasoning behind the *tf-idf* weighting is that a term occurring frequently in a document but rarely in the rest of the collection is considered to be important. Experiments have shown that *tf-idf* measure works well in many applications[17,22]. The *tf-idf* weights were calculated using a linguistic analysis tool[23].

3.4 Latent semantic analysis

LSA can be considered as an application of reduced-order Singular Value Decomposition (SVD)[23]. SVD decomposes a term-document matrix X into the product of three other matrices:

$$X = W_0 S_0 C(prime)_0 \tag{3}$$

W_0 and C_0 are the matrices of left and right singular vectors and S_0 is the diagonal matrix of singular values. W_0 has the same number of rows as the original matrix and C_0 has the same number of columns as the original matrix. S_0 is a square matrix with non-zero entries only along one central diagonal and sorted in decreasing order[1]. The dimensionality of the original matrix can be reduced by keeping the first k largest coefficients in the diagonal matrix S_0 and setting the remaining smaller ones to zero. The zero rows and columns of S_0 can then be deleted to get a new diagonal matrix S. Similarly, the corresponding columns of W_0 and C_0 can be removed to obtain W and C respectively. The product of the simplified matrices is a new matrix $\hat{X}$:

$$\hat{X} = WSC(prime) \tag{4}$$

$\hat{X}$ is the k-rank matrix with the best possible least-squares-fit to X[1]. The results of SVD include one set of k-factor loading for the terms and one for the documents. High-loading terms and documents of a factor can then be used to interpret and label the factor. For mathematical and technical details of SVD, please refer to [1](p. 397-399).

The choice of k is a critical issue in SVD. An ideal value of k should be large enough to fit all the real structure in the data and small enough to avoid unimportant details[1]. Since solutions with different number of factors represent different levels of concept aggregation, we explored 2 through 13, and 100 factors respectively. Factor interpretation and labeling was conducted manually by two MCDM researchers. The high-loading terms and documents of 2 through 13 and 100 factor solutions were examined and labeled independently. The next section discusses the results of the LSA analysis.

4 Results and discussion

4.1 Different factor solutions

This work examined 13 solutions, including 2 through 13 and 100 factors, to identify key research areas and major research themes of MCDM. For the rest of the paper, factor x-y is used to indicate the y^{th} factor of the x-factor solution[5]. For example, factor 100-2 refers to the second factor of the 100-factor solution.

Different factor solutions of LSA show different levels of research themes of the MCDM discipline. The 6-factor and 11-factor solutions describe the evolution of these areas during the past twenty-four years and reveal major research areas of MCDM, including MAUT, ELECTRE methods, analytic network process (ANP), multicriteria decision support system (MCDSS), heuristics, preference learning, interactive multiple objective programming, MCDM applications, and goal programming.

As the number of factors increases, higher level research areas can be partitioned into subareas. For example, *Preference learning* (factor 6-4 from Table 2) in the 6-factor solution is represented by *Preference representation* (factor 11-5 from Table 3) and *Preference structure modeling* (factor 11-7 from Table 3) in the 11-factor solution; and *preference modeling* (factor 100-55), *preference elicitation support* (factor 100-77), and *preference ordering techniques* (factor 100-99) in the 100-factor solution.

Table2. Top 30 High-Loading Terms for the 6-Factor Solution

Factor	Factor Label	Top 30 Terms
6-1	Analytic network process (ANP)	pro,ecis,multi,decis,multipl,pre,riteria,criteria,roc, met,problem,ultipl, gener, rel,ref,model, criterion, valu,ram,set,risk,function, experi, probabl,appli, effect,approach,object, base, altern
6-2	Multicriteria decision support system	ecis,decis,riteria,criteria,pre,met,refer,ref,valu,method, multi,base, altern, risk,maker, group,appli, experi, prefer,model,multicriteria, util, function, result,theori, regret,multipl,analysi,rel,analys

6-3	Multi-Attribute Utility Theory (MAUT)	multipl,method,pro,ultipl,decis,probabl,task,learn,met,hypothes, gener, rel,effect,criterion, fuzzi, program,experi,problem,addit,ram, approach,ecis, linear,function,pre,paper,goal,set,risk,find
6-4	Preference learning	met,multi,multipl,ultipl,decis,method,object,line,linear, process, program, pre,prefer,refer, ecis,ref,singl,learn, evalu,effici,function, ram, roc, search,task,fuzzi,solut,multiobject,paramet,probabl
6-5	ELECTRE methods	riteria,criteria,method,ecis,decis,met,valu,multi,analysi, evalu,tri, analys, object,program,ram,multicriteria, perform,solut,linear,fuzzi, multipl, select,sel,algorithm, optim,risk,line,approach,paper,develop
6-6	Heuristics	riteria,pre,multi,refer,criteria,prefer,ref,multipl,ultipl, ram,met, problem,decis, function,optim,ecis, multicriteria, object,algorithm,program,pro, fuzzi,criterion, roc,solut,system,process,tri,plan,integ

The 100-factor solution presents a large variety of research themes studied during the last twenty years by the MCDM and related disciplines (see Table 4), including MCDM theories, algorithms, related areas of research, decision support systems, applications, and techniques. It also reveals important MCDM research topics that are not presented in the 6-factor and 11-factor solutions, such as data envelopment analysis (DEA) method, genetic algorithms, simulation, behavioral issues, theoretic foundation, and visual tools.

The 100-factor solution points out two notable trends in the MCDM publications. The first is the growth in applications of MCDM. In the 100-factor solution, 21 factors are related to MCDM applications. These applications cover not only traditional application areas, such as asset management[24], scheduling problem[25], assignment problem[26], questionnaire survey[27], credit scoring[28,29,30,31], and risk evaluation[32,33,34]; but also emerging novel areas, such as verbal data classification[35], Web-based decision support[36], habitual domains[37], electronic commerce systems[38,39], and e-participation[40]. The second trend is that MCDM has entered into some new research areas[41]. For example, Supply chain management has utilized MCDM methods to capture multicriteria decision making and decision-making under uncertainty[42]. Geographical Information Systems (GIS) and MCDM have been combined to aid spatial decisions[43]. These two results generally agree with [9], [12] and [28].

Table3. Top 30 High-Loading Terms for the 11-Factor Solution

Factor	Factor Label	Top 30 Terms
11-1	Goal programming	pro,ecis,multi,decis,multipl,pre,riteria,criteria, ultipl,met,problem,roc, gener,rel,ref,model,criterion, valu,ram,set,risk,function,experi, probabl,appli, effect,approach,object,base,altern
11-2	Multiple criteria sorting problem	ecis,decis,riteria,criteria,pre,met,refer,ref, prefer,method, multi, base,altern,risk,util,group, appli,maker,experi,valu,model,multicriteria, function, result,theori,regret,multipl,analysi,criterion,rel
11-3	Interactive fuzzy multiple objective decision making	multipl,method,pro,decis,ultipl,probabl,task,learn, met,gener,ram, hypothes,rel,effect,fuzzi,program, criterion,problem,approach,experi,linear, addit, set,function,goal,risk,ecis,pre,prefer,paper
11-4	Ranking alternatives	ram,met,multi,multipl,decis,ultipl,method,line,object,process,linear, program, pre,roc,ecis,prefer,refer,ref, evalu,singl,learn,fuzzi, effici, function, search,task,solut,multiobject,analyt,valu
11-5	Preference representation	riteria,criteria,ecis,method,decis,met,valu,multi,evalu, analysi, analys,object, program,tri,solut,ram,multipl, multicriteria,perform, optim, fuzzi, select,linear, risk,sel,ultipl,line,approach,algorithm,pape
11-6	Heuristic approach	riteria,pre,multi,refer,criteria,prefer,multipl,ref,ultipl,met, ram,decis, function, problem,optim,ecis,fuzzi, multicriteria,criterion,object,algorithm,pro, process, solut,system,program,singl,risk,prioriti,roc
11-7	Preference structure modeling	met,method,riteria,ref,refer,pre,criteria,prefer,model,multipl,function, ecis,process,roc,multicriteria,ultipl, weigh,regret,algorithm,decis, weight,group,prioriti, case,gener,goal,methodolog,learn,prior,sel
11-8	Machine learning and knowledge discovery	multi,multipl,ultipl,model,system,criteria,riteria,object, valu,attribut, analys,met,search,risk,function, research,regret,method,set, analysi,man, pre,problem, effici,theori,compar,ecis,line,polici,paper
11-9	Applications	tri,riteria,criteria,attribut,met,model,method,prefer,refer,multipl,system, ultipl,problem,evalu,valu,ref, util,ram,man,multi, multicriteria, multiattribut, term, ecis,log,pro,solut,criterion,manag,analys
11-10	Multiattribute utility theory	model,multi,analys,analysi,man,ultipl,problem,object, decis,solut, multipl, program,valu,manag,ram,ecis, refer,gener,prefer,multiobject, appli,method, maker, interact,system,rel,strateg,log,algorithm,paper

11-11	Interactive procedure for MCDM	model,tri,man,valu,function,attribut,weigh,line,linear,evalu,weight, criteria, manag,problem,multi,riteria, fuzzi,util,goal,system,plan, search, cost,ram,network, research,multiattribut,decis,altern,process

Table4. Factor Labels for the 100-Factor Solution

Factor Label		
Project selection and scheduling methodology	Multiple criteria decision making under uncertainty	Outranking relations
GIS and MCDM integration	Exact algorithms	Multiple criteria linear regression
Method for ranking alternatives	Dynamic consistency (DC) optimization techniques	Monte Carlo simulation
Multi-objective optimization	Evaluating decision alternatives	Electronic commerce
MCDM in data mining	Qualitative decision making	Portfolio selection and management
ELECTRE methods	Stochastic goal programming	Multiple objective ant colony optimization algorithms
Scheduling problems	Comparative study of MCDM methods	MAUT model
Multicriteria classification	Multiple criteria simulation optimization method	Bayesian approach
Heuristic algorithm	MAVT	Preference elicitation
Interactive multiple objective programming procedure	Artificial intelligence	Interactive multiobjective optimization
Manufacturing system	Neural network for MCDM	Alternative evaluation models
Interactive multi-objective sys.	AIM	MCDM in strategic energy policy making
Multiple criteria decision support system	Environmental planning assessment and decisions	Measures of interdependences between the objectives
Design problem	Attribute weights determination	Influence diagram
Multicriteria expert support system	Tchebycheff procedure for multiple objective decision making	Multiple criteria group decision making
Genetic algorithms	Tabu search	Dynamic programming
Multi-criteria production planning	Decision maker's utility function assessment	Group decision support system (GDSS)
Information systems	Multiple criteria ABC analysis	Knowledge discovery and MCDM (neural network)
Flow shop scheduling problem	Fuzzy set and approximate reasoning	Vector optimization
Genetic algorithms	Tabu search	Dynamic programming
Algorithm development	Internet and public decision making	MCDM in cellular manufacturing system
System performance measures	Preference modeling	System design problem
Case study	Web-based decision support and applications	Game theory approach
Industrial facilities layout planning and design	ANP technique	Discrete multiple criteria problems
Operations research	Multiobjective decision making in military applications	TOPSIS
Facility location problem	Zionts-Wallenius algorithm	Graphical display tools
MCDM and industrial engineering	Modeling interaction between criteria in MCDM	Fuzzy MCDM
DSS	Applications of heuristic approaches	Lexicographic goal programming
Data mining and ML	AHP improvements	DEA
Philosophy of MCDM	Parameter determination methods	Theoretic foundation
Goal programming	Metaheuristic algorithm	Behavioral issues
Multicriteria location problem	SMAA	Optimization algorithms and implementation of MCDM
Resource allocation model	Team decision making under uncertainty	
Visual tools	Simulation modeling	

Table5. Top 10 High-Loading Papers for the 6-Factor Solution

Factor	High-Loading Papers	Factor Loading
6-1	Jin Woo Lee, Soung Hie Kim, C&OR,2000	0.14
	J. M. Coutinho et al.,C&OR ,1999	0.11
	Wey, Wann-Ming,Wu, Kuei-Yang, M&CM,2007	0.11
	Behnam Malakooti, Jumah E. Al-alwani, C&OR, 2002	0.09
	Minghe Sun et al., C&OR, 2000	0.08
	Lorraine R. Gardiner, Ralph E. Steuer, EJOR, 1994	0.08
	Otto Rentz, FSS, 1996	0.08
	Taeyong Yang et al., FSS, 1991	0.08
	Bernard Roy, Roman Slowinski, AOR, 2006	0.07
	Mark A. Coffin, Bernard W. Taylor, C&OR, 1996	0.07
6-2	C. Zopounidis, Michael Doumpos, C&OR, 2000	0.14
	T. Terlaky, EJOR, 1985	0.11
	V. Mousseau et al., C&OR, 2000	0.11
	Taeyong Yang et al., FSS, 1991	0.11
	Otto Rentz, FSS, 1996	0.11
	Lorraine R. Gardiner, Ralph E. Steuer, EJOR, 1994	0.09
	N. M. Badra, FSS, 2002	0.09
	E. Melachrinoudis, Z. Xanthopulos, C&OR, 2003	0.09
	Masatoshi Sakawa, Hitoshi Yano, FSS, 1989	0.09
	John A. Aloysius, et al., EJOR, 2006	0.08
6-3	Jose Rui Figueira et al., EJOR, 2008	0.15
	Risto Lahdelma et al., EJOR, 2003	0.15
	David L. Olson, EJOR, 2001	0.15
	S. Greco, V. Mousseau, R. Slowinski, EJOR, 2008	0.11
	Stelios H. Zanakis, et al., EJOR, 1998	0.1
	Pekka J. Korhonen, Jukka Laakso, EJOR, 1986	0.1
	Gerard Colson, C&OR, 2000	0.09
	Silvia Angilella, EJOR, 2004	0.09
	An Ngo The, Vincent Mousseau, JMD,2002	0.09
	Risto Lahdelma, Pekka Salminen, EJOR, 2002	0.08
	Edmund Kieran Burke, Sanja Petrovic, EJOR, 2002	0.08
6-4	Jose Rui Figueira et al.,EJOR,2008	0.25
	Bernard Roy, Roman Slowinski, EJOR, 2008	0.19
	George Mavrotas, Panagiotis Trifillis, C&OR, 2006	0.19
	Theodor J. Stewart, EJOR, 1986	0.18
	Peter Muller, DA, 2006	0.13
	Kim Fung Lam, Eng Ung Choo, JORS, 1995	0.1
	Murat Koksalan, Ahmet Burak Keha, 2003	0.1
	Risto Lahdelma, Pekka Salminen, EJOR, 2002	0.09
	Salvatore Greco, et al., EJOR, 2002	0.09
	Gregory E. Kersten, DSS, 1988	0.08
6-5	J.C. Leyva-Lopez, E. Fernandez-Gonzalez, EJOR, 2003	0.28
	Salvatore Greco et al., EJOR, 2008	0.25
	Bernard Roy, Roman Slowinski, EJOR, 2008	0.12
	Risto Lahdelma, Pekka Salminen, EJOR, 2002	0.12
	J OS C. FODOR, MARC ROUBENS, JMCDA, 1997	0.12
	Silvia Angilella et al., EJOR, 2004	0.11
	Huseyin Cavusoglu, Srinivasan Raghunathan, DA, 2004	0.11
	Salvatore Greco, et al., EJOR, 2002	0.1
	Minghe Sun, EJOR, 2002	0.1
	Minghe Sun, EJOR, 2002	0.1
6-6	J. Gupta, Kruger, Lauff, Werner, Sotskov, C&OR, 2002	0.27
	J. Gupta, K. Hennig, F. Werner, C&OR, 2002	0.24
	Peter Muller et al., DA,2006	0.13
	Sandeep Purao et al., DSS, 1999	0.12
	Gregory E. Kersten, DSS, 1988	0.12
	Ilia Tsetlin, Robert L. Winkler, DA, 2006	0.12
	Jatinder N. D. Gupta, Johnny C. Ho, C&OR, 2001	0.11
	Vincent T'kindt et al., C&OR, 2003	0.11
	B. Malakooti, C&OR, 1989	0.1
	Julian Molina et al., EJOR, 2008	0.1

4.2 Different factor solutions

Table6. Factor Labels and Paper Counts for the 6-Factor Solution

Factor	Factor Label	Paper Counts				
		85-89	90-94	95-99	00-04	05-09
6-1	Analytic network process (ANP)	41	48	54	63	61
6-2	Multicriteria decision support system	37	63	51	49	39
6-3	Multi-Attribute Utility Theory (MAUT)	17	44	62	40	42
6-4	Preference learning	22	27	34	22	55
6-5	ELECTRE methods	38	42	54	59	35
6-6	Heuristics	32	36	49	39	45

Table7. Factor Labels and Paper Counts for the 11-Factor Solution

Factor	Factor Label	Paper Counts				
		85-89	90-94	95-99	00-04	05-09
11-1	Goal programming	21	20	37	27	16
11-2	Multiple criteria sorting problem	37	63	51	49	39
11-3	Interactive fuzzy multiple objective decision making	15	21	13	19	11
11-4	Ranking alternatives	22	14	26	24	7
11-5	Preference representation	21	16	29	34	30
11-6	Heuristic approach	24	22	27	39	29
11-7	Preference structure modeling	19	22	37	26	33
11-8	Machine learning and knowledge discovery	14	16	25	29	16
11-9	Applications	27	29	45	32	26
11-10	Multiattribute utility theory	17	19	19	18	20
11-11	Interactive procedure for MCDM	18	24	37	28	35

Figure 2 suggests that the growth in some research areas, such as preference learning (factor 6-4), heuristics (factor 6-6), and analytic network process (ANP) (factor 6-1) increased considerably from the 1985-1989 period to the 2005-2009 period. In the case of ELECTRE methods (factor 6-5), the number of publications maintained a relatively constant increase from 1985-2008. The research interests in Multi-Attribute Utility Theory (factor 6-3) grew significantly from 1985-1989 to 1995-1999 and dropped during the 2000-2004 period. The number of publications in MAUT remained stable since then. Multicriteria decision support system (factor 6-2) experienced a rapid growth from 1985-1989 to 1990-1995 and declined slightly during 1995-2008.

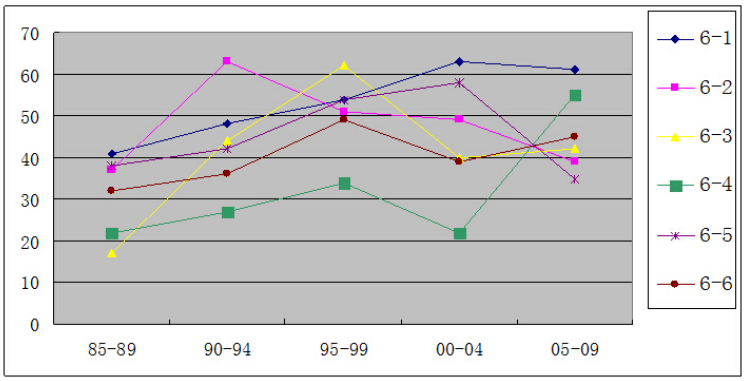

Figure 2: Dynamics of Major Research Areas (six-factor solution)

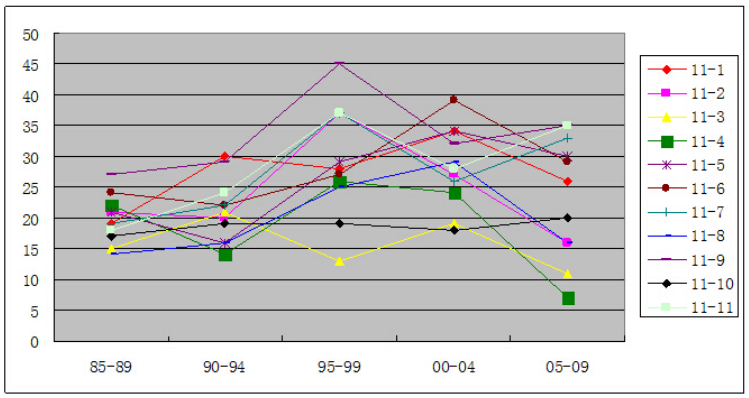

Figure 3: Dynamics of Major Research Areas (eleven-factor solution)

5 Conclusions and limitations

This paper attempted to identify the major research areas and themes of MCDM field by examining a large body of related research papers using latent semantic analysis. In the experimental study, over fifteen hundred abstracts of MCDM/MAUT field were collected and analyzed to obtain thirteen factor solutions. The 6-factor and 11-factor solutions of the analysis reveal key research areas of MCDM/MAUT. MAUT, ELECTRE methods, ANP, multicriteria decision support system (MCDSS), heuristics, preference learning, interactive multiple objective programming, MCDM applications, and goal programming are among the main streams of thought of the field.

The ideas and techniques of MCDM are continuing to integrate into other disciplines. For example, data mining (DM) field used ELECTRE methods to cluster opinions[44] and utilized multiple criteria decision aid process to help users to sort association rules[45]. Artificial neural networks, an artificial intelligence (AI) method, has been used by MCDM researchers to solve discrete MCDM problem[46] and model decision-makers' preference structures[47]. Geographical Information Systems (GIS) and MCDM have been combined to aid spatial decisions[43].

This study has several limitations. First, since the LSA analysis depends on identifying frequent word usage patterns from a collection of text, it is difficult to capture a research area if it is not well established and has not established consistent terminology among its researchers[5] . Second, this study only collected articles published after 1985 because the major areas and the evolution of MCDM and MAUT before 1990s have been investigated in previous studies[10,11,12]. Third, the research abstracts collected in this analysis include only English language journals. Papers published in other languages are not considered.

Acknowledgements

This work was supported in part by grants from the National Natural Science Foundation of China (#71325001, #71222108 and #71173028), Program for New Century Excellent Talents in University (NCET-12-0086).

Bibliography

[1] Deerwester, S.; Dumais, S.; Furnas, G.; et al. (1990). Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, 41(6): 391-407.

[2] Landauer, T.; Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge, *Psychological Review*, 104: 211-240.

[3] Landauer, T.; Foltz, P.; Laham, D. (1998). Introduction to Latent Semantic Analysis, *Discourse Processes*, 25: 259-284.

[4] Kou, G.; Lou, C. (2012). Multiple Factor Hierarchical Clustering Algorithm for Large Scale Web Page and Search Engine Clickstream Data, *Annals of Operations Research*, 197(1)25: 123-134.

[5] Sidorova, A.; Evangelopoulos, N.; Valacich, J. S.; et al. (2008). Uncovering the intellectual core of the information systems discipline, *MIS Quarterly*, 32(3): 467-482.

[6] Dumais, S. T.; Furnas, G. W.; Landauer, T. K.;et al (1988). Using latent semantic analysis to improve information retrieval, *Proceedings of CHIŻ88 Conference on Human Factors in Computing Systems*, 281-285.

[7] Dumais, S. T. (2004). Latent Semantic Analysis, *Annual Review of Information Science and Technology*, 38: 189-230.

[8] Gansterer, W.N.; Janecek, A.G.K.; Neumayer, R. (2008). In M. W. Berry and M. Castellanos (eds.), Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition (pp. 165-183). Springer

[9] Wallenius, J.; Dyer, J. S.; Fisburn, P. C.; et al. (2008). Multiple Criteria Decision Making, Multiattribute Utility Theory: Recent Accomplishments and What Lies Ahead, *Management Science*, 54(7): 1336-1349.

[10] Stewart T. J. (1992). A critical survey on the status of multiple criteria decision making theory and practice, *OMEGA*, 20(5/6): 569-586.

[11] Dyer, J. S.; Fisburn, P. C.; Steuer, R. E.; et al. (1992). Multiple criteria decision making, multiattribute utility theory: the next ten years, *Management Science*, 38(5): 645¨C654.

[12] Urli, B.; Nadeau, R. (1999). Evolution of multi-criteria analysis: a scientometric analysis, *J. Multi-Crit. Decis. Anal.*, 8: 31-43.

[13] International society on MCDM, (2009). http://www.mcdmsociety.org, Accessed 24 Jun 2009

[14] Steuer, R. E.; Gardiner, L. R.; Gray, J. (1996). A bibliographical survey of the activities and international nature of multiple criteria decision making, *J. Multi-Crit. Decis. Anal.*, 5: 195¨C217.

[15] Bragge, J.; Korhonen, P.; Wallenius, J.; et al. (2008). Bibliometric Analysis of Multiple Criteria Decision Making/Multiattribute Utility Theory, *International Society on Multiple Criteria Decision Making*, Accessed 11 June 2009.

[16] Fox, C. (1992). Lexical Analysis and Stoplists. In W. B. Frakes and R. Baeza-Yates (eds.), Information Retrieval: Data Structures and Algorithms (pp. 102-130). Upper Saddle River, NJ: Prentice-Hall.

[17] Han, J.; Kamber, M. (2006). Data Mining: Concepts and Techniques, 2nd edition. San Francisco, CA: Morgan Kaufmann Publishers.

[18] Stopwords. (2008). Webconfs.com, `http://www.webconfs.com/stop-words.php`, Accessed 10 August, 2008.

[19] SQL Sever 2005. Microsoft.com, `http://www.microsoft.com/sqlserver/2005/en/us/overview.aspx`, Accessed 1 Feb 2009.

[20] Porter, M. F. (1980). An algorithm for suffix stripping, *Program*, 14(3): 130-137.

[21] Porter, M. F. (2008). The Porter Stemming Algorithm. http://tartarus.org/ martin/ Porter-Stemmer/. Accessed 22 Feb, 2009.

[22] Baeza-Yates, R.; Ribeiro-Neto, B. (1999). Modern Information Retrieval, *Addison-Wesley, Wokingham*, UK.

[23] LingPipe (2008). `http://alias-i.com/lingpipe/index.html`, Accessed 1 March 2009.

[24] Langen, D. (1989). An (interactive) decision support system for bank asset liability management,*Decision Support Systems*, 5(4): 389-401.

[25] Geiger, M. J. (2007). On operators and search space topology in multi-objective flow shop scheduling, *European Journal of Operational Research*, 181(1): 195-206.

[26] Przybylski, A.; Gandibleux, X.; Ehrgott, M. (2008). Two phase algorithms for the bi-objective assignment problem. *European Journal of Operational Research*, 185(2): 509-533.

[27] Ergu, D.; Kou, G. (2012). Questionnaire Design Improvement and Missing Item Scores Estimation for Rapid and Efficient Decision Making, *Annals of Operations Research*, 197(1):5¨C23, DOI 10.1007/s10479-011-0922-3.

[28] Shi, Y. (2001). Multiple Criteria Multiple Constraint-level (MC2) Linear Programming: Concepts, Techniques and Applications, *World Scientific Publishing*, 539 pages.

[29] Yu, L.; Wang, S.; Lai, K. K. (2009). An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring. *European Journal of Operational Research*, 195(3): 942-959.

[30] Kou, G.; Peng, Y.; Wang, G.X. (2014a). Evaluation of Clustering Algorithms for Financial Risk Analysis using MCDM Methods, *Information Sciences*, 27:1-12, DOI: HTTP://DX.DOI.ORG/10.1016/j.ins.2014.02.137.

[31] Kou, G.; Peng, Y.; Lu, C. (2014b). An MCDM Approach to Evaluate Bank Loan Default Models, *Technological and Economic Development of Economy*, 20(2): 278¨C297, DOI: HTTP://DX.DOI.ORG/10.3846/20294913.2014.913275.

[32] Ergu, D.; Kou, G.; Shi, Y.; et al. (2011). Analytic Network Process in Risk Assessment and Decision Analysis, *Computers & Operations Research*, DOI: 10.1016/j.cor.2011.03.005.

[33] Kou, G.; and Lin, C. (2014) A cosine maximization method for the priority vector derivation in AHP, *European Journal of Operational Research*, 235: 225-232 , DOI: http://DX.DOI.ORG/10.1016/j.ejor.2013.10.019

[34] Montibeller, G.; Belton, V.; Lima, M.V.A. (2007). Supporting factoring transactions in Brazil using reasoning maps: a language-based DSS for evaluating accounts receivable. *Decision Support Systems*, 42(4): 2085-2092.

[35] Yevseyeva, I.; Miettinen, K.; Rasanen, P. (2008). Verbal ordinal classification with multicriteria decision aiding. *European Journal of Operational Research*, 185(3): 964-983.

[36] Hamalainen, R. P. (2003). Decisionarium-aiding decisions, negotiating and collecting opinions on the web. *Journal of Multicriteria Decision Analysis*, 12(2-3): 101-110.

[37] Yu, P.L. (1991). Habitual domains, *Operations Research*, 39(6): 869-876.

[38] Chiu, Y.; Shyu, J. Z.; Tzeng, G. H. (2004). Fuzzy MCDM for evaluating the e-commerce strategy, *International Journal of Computer Applications in Technology*, 19(1): 12-22.

[39] Kameshwaran, S.; Narahari, Y.; Rosa, C. H.; et al. (2007). Multiattribute electronic procurement using goal programming. *European Journal of Operational Research*, 179(2): 518-536.

[40] Moreno-Jimenez, J. M.; Polasek, W. (2003). e-democracy and knowledge. A multicriteria framework for the new democratic era. *Journal of Multi-Criteria Decision Analysis*, 12(2-3): 163-176.

[41] Zeleny, M. (1998). Multiple criteria decision making: eight concepts of optimality, *Human Systems Management*, 17(2): 97-107.

[42] Dong, J.; Zhang, D.; Yan, H.; et al. (2005). Multitiered Supply Chain Networks: Multicriteria Decision Making Under Uncertainty. *Annals of Operations Research*, 135(1): 155-178.

[43] Gomes, E. G.; Lins, M. (2002). Integrating geographical information systems and multicriteria methods: A case study. *Annals of Operations Research*, 116(1-4): 243-269.

[44] Bisdorff, R. (2002); Electre-like clustering from a pairwise fuzzy proximity index, *European Journal of Operational Research*, 138(2): 320-331.

[45] Lenca, P.; Meyer, P.; Vaillant, B.; et al. (2008). On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, 184(2): 610-626.

[46] Malakooti, B.; Zhou, Y. Q. (1994). Feedforward Artificial Neural Networks for Solving Discrete Multiple Criteria Decision Making Problems, *Management Science*, 40(11): 1542-1561.

[47] Wang, J. (1994). A neural network approach to modeling fuzzy preference relations for multiple criteria decision making. *Computers and Operations Research*, 21(9): 991-1000.