Research Trend Analysis using Word Similarities and Clusters

KyoJoong Oh, Chae-Gyun Lim, Sung Suk Kim and Ho-Jin Choi

Department of Computer Science, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Korea
{aomaru, rayote, powerkimss, hojinc}@kaist.ac.kr

Abstract

In this paper, we propose a new research trend analysis using important word clusters and its relationship. Journals published many papers every month or week and new scientific contributions were exponentially cumulated to their database. If can analysis important words and related relationships of the papers, a change of research trend in a domain is an interesting topic in text mining. We use a Term Frequency Inverse Document Frequency (TFIDF) to extract meaningful words, the similarity of words measures using WordNet information and a document comparison approach. To measure the similarity from word lists extracted by TFIDF and differences of important word clusters and weights, the approach analyzes the research trend and visualizes the differences of research interest in same research fields. To show usefulness of proposed approach, we illustrate simulations and various results.

Keywords: *TFIDF, Word Similarity, Important Word Cluster, Research Trend, Word Network Graph, Document Comparison*

1. Introduction

Today, information had rapidly increased through developing communication abilities such as information retrieval and accessibility to the Internet. Scholarly documents including scientific papers were also cumulated in online digital archives consisted of text representation [1]. Therefore text mining is an indispensable process in analysis of the documents [2-8]. Human understands and reading abilities rare improved which relatively compared with an increase of the amount of information. Information acquisition of the human's interesting is one of the essential processes using the text mining in huge information [5, 7, 9]. Text mining has served users to useful information from the online digital archives which published every weeks or months [2-9]. For example, a Term Frequency Inverse Document Frequency (TFIDF) was a fast tool to extract useful words as the value of the information from documents [9].

We proposed a new approach to find and trace research trends using the text mining. The approach, in first step, extracts meaningful words from the papers separated by the year using the TFIDF. Next step is a generating important word clusters consisted of unique words. Using the word clusters, the approach estimates the similarity among the words for each year and all years. We verify the research trends using comparison results of the similarity. In addition, we can also represent a word similarity network for easy-to-understand relationships and distributions of words.

In Section 2, we described related methods which include the TFIDF and similarity measurement. A proposed approach is explained in Section 3. Then, we prove usefulness of proposal in Section 4. Finally, we summary and conclude in Section 5.

2. Methods

We discuss related methods including existed techniques and proposed measurements to illustrate the proposed approach. We shortly introduce a Term Frequency Inverse Document Frequency (TFIDF) [9] which is a well-known tool of the text mining techniques to extract meaningful words from the documents.

The Term Frequency (TF) counts how many frequent each term appears in a document, as following equation.

$$tf(t,d) = count(t \in d) \qquad \textbf{(1)}$$

And, Document Frequency (DF) counts how many frequent each term appears in documents, as following equation.

$$df(t,D) = |\{d|t \in d, d \in D\}| \qquad \textbf{(2)}$$

Inverse Document Frequent (IDF) is the inverse of the DF. IDF checks how many documents have a specific term, and assumes that the term is less useful when appears in many documents, as following equation

$$idf(t,D) = log\frac{|D|}{df(t,D)} \qquad \textbf{(3)}$$

Finally, The TFIDF is computed by the combination of the TF and the IDF as follows

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \qquad \textbf{(4)}$$

where TF $\mathrm{tf(t,d)}$ is counted number in a document, IDF $idf(t,D)$ is log of $|D|/df(t,D)$ and DF $df(t,D)$ is already explained. After the TFIDF algorithm, we can obtain meaningful words from the documents and sets of the documents.

To find a connection of meaning between the words, we use the word similarity tool based on [11, 12] and word dictionary such as WordNet [13]. Using the WordNet information and relevant similarity measurements, the word similarity connections and weights are obtained for generating word network graph to more easily understand a structure of word distributions.

$$sim_k = \sum_{i=1}^{n} exp^{-\left(\frac{(w_i - c_i)^2}{\sigma^2}\right)} \qquad \textbf{(5)}$$

where w_i is the ith word weight of the specific year TFIDF and c_i is same word weight of the important word clusters (IWC) such as unique words. It is also applied to other similarity measurements which include interconnection word similarities and the research trend analysis.

The IWCs and interconnect words are important concepts in this paper. The IWCs are extracted from the TFIDF of each year and delete duplicated terms that consist of the words and relevant weights. Then, a next process finds a uniqueness of the words that are didn't find the documents of other years. The unique words in the IWC have high probabilities of non-duplications between other years. A member of the cluster has two entities including word and related weight. We assume that these word clusters can representative sets of the year.

The interconnection words are extracted by connections of the word similarity between two years in whole TFIDF words. An interconnection word has both-side similarity relationships between previous year and the following year.

3. Proposed Approach

We describe following sequential steps to extract IWCs and interconnection words to represent the research trend through numerical changes in tables and graphical distributions.

Step 1. Separation of Document

We select a domain in a journal. We collect whole papers of the domain. We manually separate the documents by the year. It means that we want to analyze the research trends by year. In this process, we remove stop words using general stop words lists on the web, and we do stemming using WordNet library.

Step 2. Extraction of Meaningful Words

We extract the meaningful words by each paper and year using the TFIDF. We select 50 top-ranked papers according to the numbers of citations by years. In next, we extract 50 numbers of the meaningful words by each paper. After that, we rearrange the meaningful words by the year without redundant words. Then we can get the meaningful words by year.

Step 3. Generation of Uniqueness

We generate the unique words lists called by the uniqueness from the meaningful words by year. We call it important word cluster (IWC) which contains adjusted weights. The important word cluster represents the research trends by year in specific domain using the words distribution. The figure shows the words distribution where the clusters are located in the all words distributions.

Step 4. Measurement of Word Similarity

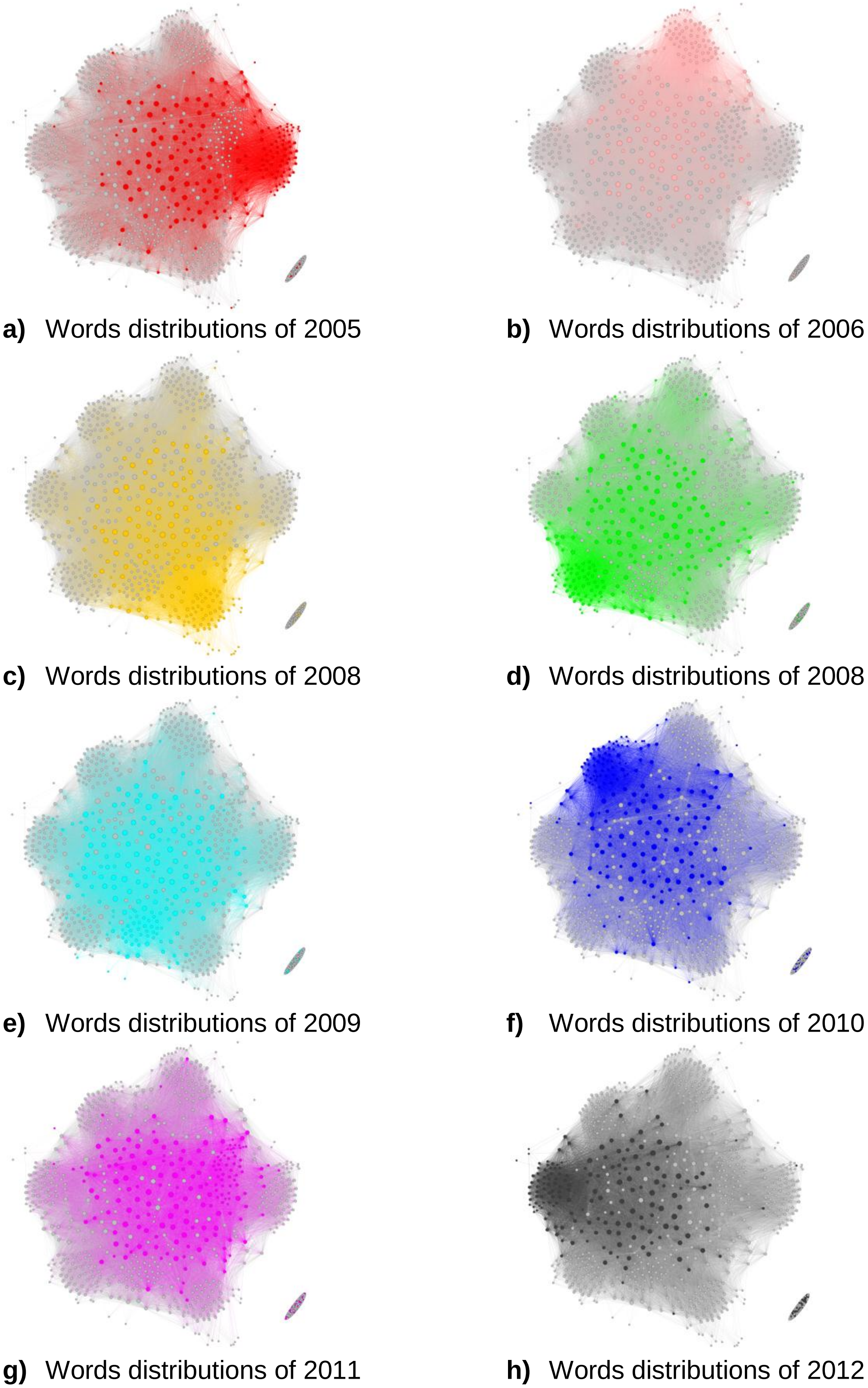
We measure the similarity among the unique words by year and verify the research trend using the IWCs. In addition, we also measure the similarity among the meaningful words of all papers to figure out the word similarity network.

Step 5. Extraction of Interconnection Words

We also find the interconnection words connected between the important word clusters in previous year and the following year. We can find in-directed changes of the research trends by year. The figure shows the changes of the research trends in the specific domain year by year using the interconnection words and the distribution of unique words.

Step 6. Measurement of Similarity of Interconnection Words

The IWCs are generated from the meaningful words of the TFIDF which have higher weights in the periods. It means that IWC words have higher TF values although there are unique words. Such characteristics can enhance the analysis of the changing trend and representative words.



a) Words distributions of 2005

b) Words distributions of 2006

c) Words distributions of 2008

d) Words distributions of 2008

e) Words distributions of 2009

f) Words distributions of 2010

g) Words distributions of 2011

h) Words distributions of 2012

Figure 1. Words Distributions by Year from TFIDF

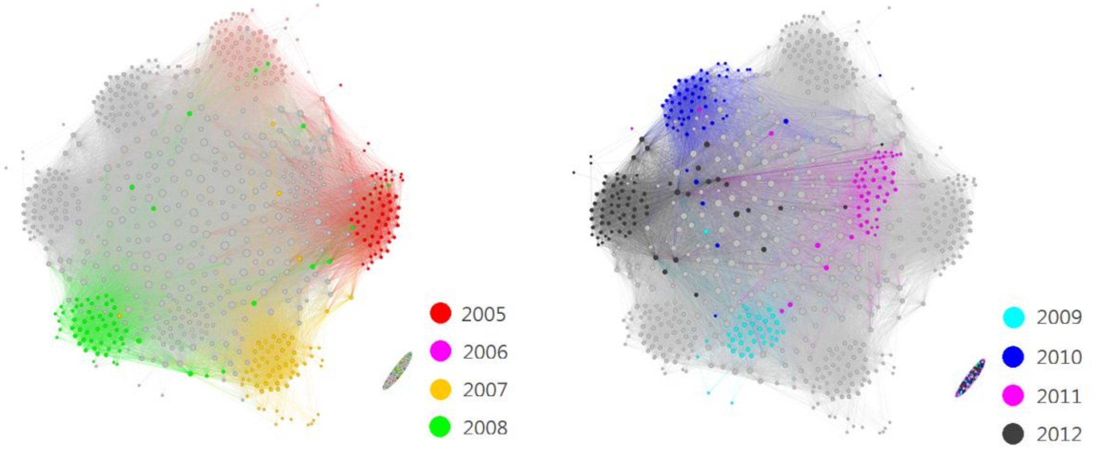
4. Simulation and Results

We used the document sets that consist of Bioinformatics journal from 2005 to July 2012 obtained published papers. The papers were used to learn the research trends by years. We visited the site of Bioinformatics journal and gathered the papers according to tables of contents page. We implemented a program that did these processes automatically. After that, we filtered stop words and did stemming using WordNet library as preprocessing, also we implemented the autonomous program to do the preprocessing.

Figure 1 shows all words network of the specific domain in this case “Genetic Analysis”. There are 867 numbers of word nodes and 45553 numbers of edges in all words network. We selected 200 top-ranked words for each year and merged same words remained similarity relationships. Red nodes in a) are the IWC of 2005, pink nodes in b) are from 2006, yellow nodes in c) are from 2007, green nodes in d) are from 2008, cyan nodes in e) are from 2009, blue nodes in f) are from 2010, magenta nodes in g) are from 2011, and dark gray nodes in h) are from 2012 respectively. Departed nodes in right bottom didn’t find the similarity relationships from the WordNet 2.1 dictionary. Using this figure, we can distinguish the distributions and changes of the IWCs.

An interesting points in Figure 1 is the words distributions of 2009 e) and 2011 g) are shifted to center compared with other years. It means that the IWCs of these two years have many words relationships with other years. The papers accepted by the journal in 2009 and 2011 used highly relevant words with other years. In other words, the terminologies of a particular domain are used commonly as time passes.

The commonly used words were included in Figure 1 by each year. Therefore, we figured out another graph without the common words in Figure 2. We could distinguish the words clusters and find differences from the words distributions by year more easily. The words lists would be the IWCs by year.



a) Words distributions 2005 to 2008 **b)** Words distributions 2009 to 2012

Figure 2. Words Distributions without Common Words by Year from TFIDF

Table 1 contains some lists of important words clusters. We calculated values of weights based on similarity in TFIDF and WordNet Similarity. The words lists represent the research trends by year.

Table 1. Results of IWCs by Year

2005	2006	2007	2008	2009	2010	2011	2012
ion	histone	chisel	motif	motif	therapy	alignment	phage
mass	acetylation	primer	evigan	anchor	bin	nucleosome	locus
vertex	gpt	genomix	assembly	predictor	match	screen	gms
longsage	mask	genotype	cabog	phage	triangle	pyicos	lifestyle
tfbs	genotype	hairpin	allele	items	snp	motif	somatic
splice	tss	sap	optical	utgb	event	score	nucleosome
operon	nonlinear	microprocessor	ligand	ddi	silencer	assembly	rare
seed	motif	minority	genotype	cytogenetic	lcp	deletion	edcf
sis	asti	unlabeled	mantis	tumor	acs	conveyor	indel
network	cwt	gpas	branch	suffix	segmentation	svseq	replicate
mismatch	array	attribute	principal	classifier	array	pindel	crm
annotation	tandem	svm	kernel	snp	marker	probabilistic	genotype
gps	ftr	hypermethylated	artemis	locus	enzyme	breakpoints	sybil
genome	solution	contigs	megablast	category	chromatin	reducer	pfam
clone	tissue	layout	fdr	alignment	egm	bind	cnv
graph	ucsc	hit	vector	lysogenic	ctgdr	block	splazers
agml	rlmm	phylogenetic	index	gada	cog	cycle	microbiome
translocation	wavelet	signature	intensity	endophenotypes	permutation	plate	fsr
pfsm	quartet	prediction	edit	array	rule	snp	artemis
copy	statistic	substitution	mitochondrial	hairpin	phenotype	node	contigs
cytokine	fdr	band	normalization	prediction	predictive	eta	strelka
feature	hgt	array	rna	pairagon	assembly	flash	segment
oligos	vamp	enzymatic	query	msmad	nucleosome	trait	gwas
edge	greedy	bind	italics	browser	drug	marker	epistasis
dfa	bind	tree	smap	toolkit	cmds	column	path
intron	period	literal	primer	endophenotype	pathgroups	segment	trait
dispensability	score	microsatellites	pathway	tile	alignment	nod	glycosylation
transcript	snp	logic	stem	potency	dcj	pscbs	allele
interaction	mixture	blast	structure	methylation	concept	phenotype	snp
chromosome	unlabeled	site	consensus	taxonomic	snpruler	prediction	mappability
codon	ivom	oslay	isoform	finder	similarity	pathway	fragment
sample	transcription	snp	ancestral	primer	nova	games	lpchp
loop	activity	contig	rnatops	site	gap5	site	indels
ndfa	negative	mams	hierarchy	smooth	recurrent	bambus	region
model	alignment	glimmer	breakpoints	chip	penalty	target	mdr
halve	rf2	disease	trio	megan	satsuma	normalization	tumor
selenoprotein	chip	assembly	variant	hhmm	weight	insert	score
oligo	target	metabolic	laser	giddi	lasso	sis	motif
peptide	haplotype	affinity	aimie	intensity	object	transcript	assembler
promoter	profile	alignment	family	interaction	enhancer	annotation	variant
gene	gene	sequence	state	copy	annotation	edge	insertion
protein	tag	undetermined	sequence	seed	edge	copy	abundance
probe	protein	mass	specie	feature	gps	sample	classification
tag	probe	interaction	tree	probe	sample	graph	null
profile	promoter	annotation	clone	gene	interaction	seed	codon
exon	cluster	exon	copy	profile	probe	model	interaction
cluster	train	cluster	gene	cluster	gene	gene	sample
train	exon	train	probe	train	protein	profile	feature
peak	peak	peak	peak	peak	cluster	exon	promoter
spectrum	spectrum	spectrum	spectrum	spectrum	train	peak	protein

We calculated similarities between the IWCs and the 50 top-ranked words for each paper, and also calculated the average value by year.

The variance of similarity function σ is 100 and the values of sim_k were amplified up to 10 times to adjust.

Table 2. Similarity between IWCs and 50 Top-ranked Words

	2005	2006	2007	2008	2009	2010	2011	2012
2005	**41.763**	28.82928	32.30753	28.48492	24.80209	25.02819	25.17764	21.699
2006	23.39757	**41.6279**	28.50047	30.18784	28.60414	27.49522	27.19125	22.52223
2007	22.58571	21.5334	**46.0276**	26.52565	22.14796	24.88311	22.77369	21.01121
2008	17.64583	20.76893	23.07059	**47.6118**	23.33962	20.60183	21.55708	20.15477
2009	21.60992	28.69268	31.14828	31.09755	**36.6449**	28.22292	25.13034	20.15832
2010	18.78853	20.55041	25.51655	24.47164	23.35072	**40.8712**	23.98109	24.90493
2011	27.98659	24.08971	31.63655	31.73786	28.29755	27.72717	**39.1681**	28.11638
2012	12.23995	20.13594	18.3422	20.65713	21.40456	23.63179	25.11879	**42.961**

Table 2 is a table of results comparison of the similarities which consists of the numerical values between the IWCs and the TFIDFs of each year. We confirm that the similarity in the same year to high similarity TFIDFs and IWCs, but the other interlocking. Additionally, it can be found numerically that the similarity is lower farther away the time. Figure 3 shows the typical results as graph.

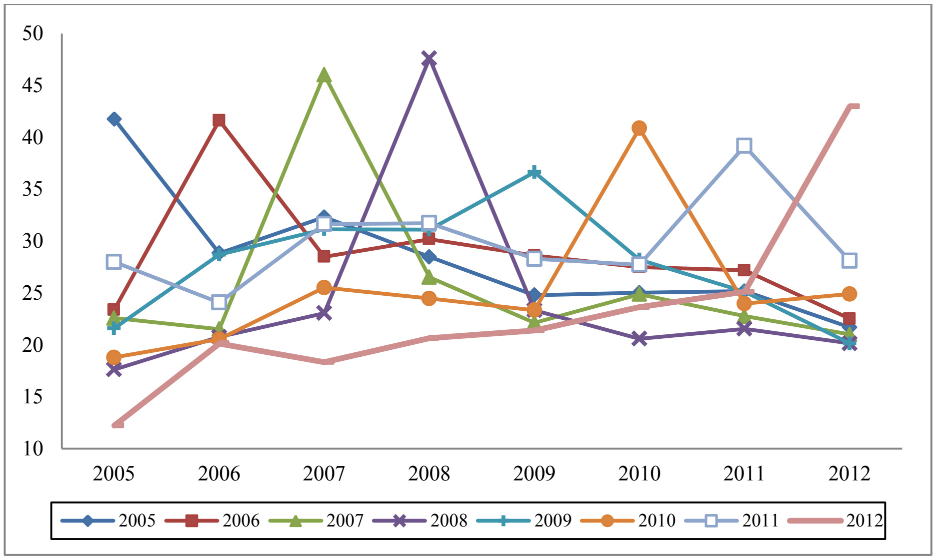


Figure 3. Similarity Changes between IWCs and 50 Top-ranked Words

In addition, we found lists of interconnection words which had relationships between IWCs over two years. We calculated the similarities between the interconnection words and the 50 top-ranked words for each paper, and also calculated the average value by year.

The variance of similarity function σ is 100 and the values of sim_k were amplified up to 10 times to adjust same as above.

Table 3. Similarity between Interconnection Words and 50 Top-ranked Words

	2005	2006	2007	2008	2009	2010	2011	2012
05-06	2.433852	2.821333	2.904639	2.633269	2.14282	2.307651	2.039968	1.879988
06-07	2.339542	2.714175	2.999853	2.966597	2.28568	2.4615	2.499958	2.039979
07-08	2.094264	2.464195	2.99992	3.299917	2.02377	2.794832	2.679961	2.759973
08-09	2.131966	2.392777	2.714232	3.399856	2.357072	2.384584	2.419954	2.319977
09-10	2.226358	2.714247	3.476117	3.333284	2.690425	2.87175	2.719975	2.99997
10-11	2.754658	3.071381	3.666622	3.833287	2.928535	3.794814	3.439916	3.399961
11-12	1.660361	1.892845	2.428554	2.266609	2.071384	2.102532	2.159944	2.439952

Table 3 is a table of results comparison that measures the similarities of the interconnection words and 50 top-ranked words for each paper. The result shows average values of the similarities which are obtained between the interconnection words and 50 top-ranked words of each paper. The interconnection words are selected by the WordNet similarity more than 0.2. Figure 4 shows the results as graph. The similarities between interconnection words and 50 top-ranked words were relatively high over related two years.

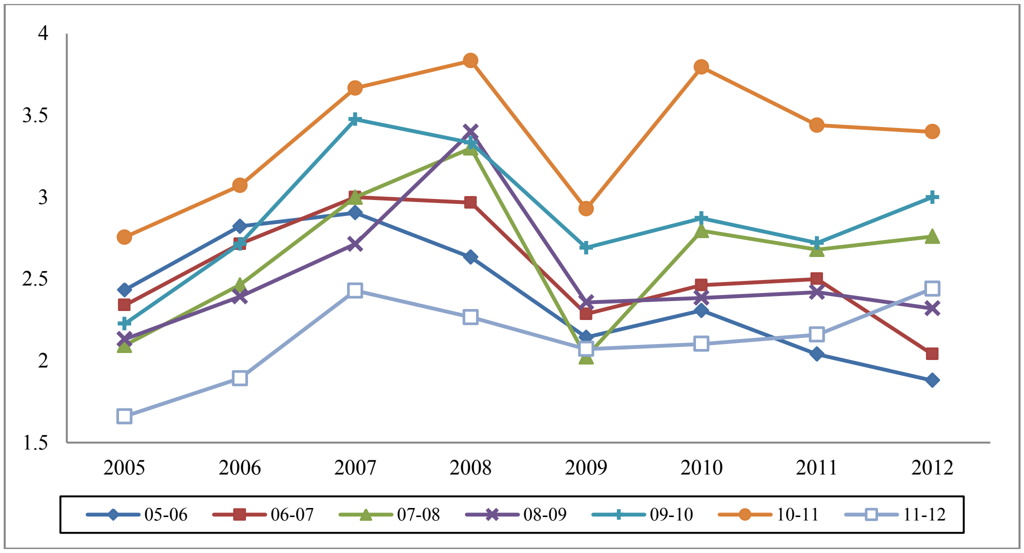


Figure 4. Changes of Similarity between Interconnection and 50 Top-ranked Words

At last, we compared each other lists of interconnection words by year. We counted numbers of words commonly used.

Table 4. Numbers of common words from each other lists of interconnection words

	05-06	06-07	07-08	08-09	09-10	10-11	11-12
05-06	**48**	33	21	21	27	29	13
06-07	33	**52**	32	23	31	35	20
07-08	21	32	**60**	34	28	33	18
08-09	21	23	34	**52**	37	28	14
09-10	27	31	28	37	**55**	39	18
10-11	29	35	33	28	39	**72**	34
11-12	13	20	18	14	18	34	**49**

Table 4 shows a table of results including numbers of common words from each other lists of the interconnection words. It means that a list of the interconnection words is how similar to other lists. The bold values are numbers of all words in each list. We can notify that the similarity also reduced from gap of numerical values according to time passed in Figure 5.

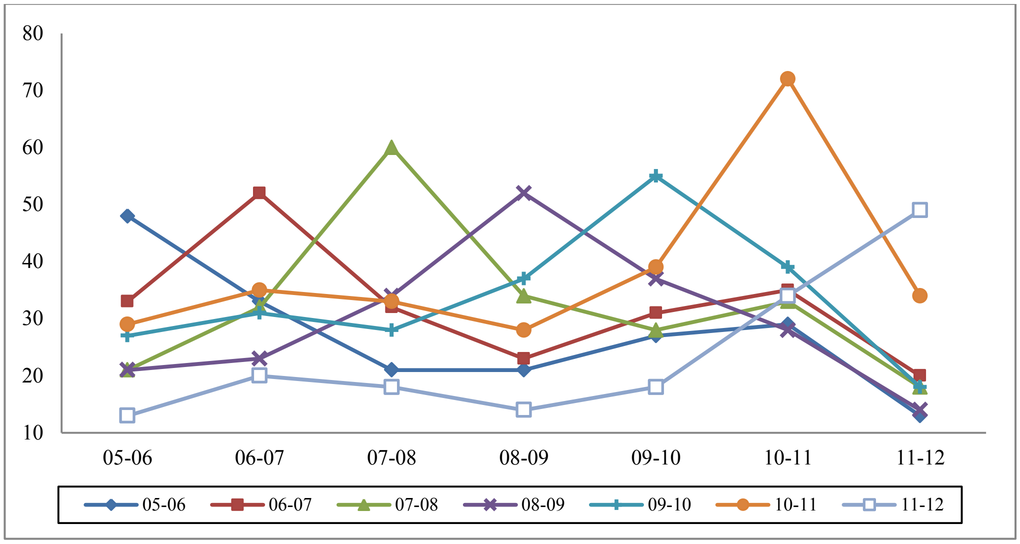


Figure 5. Changes of Numbers among Interconnected Words

As shown in the figures and the tables, we confirmed the research trend using the IWCs and the interconnection words. Each IWC had a uniqueness that compared with other years through the changes of numerical values. The words of the IWCs can be considered by the representing words by year. Additionally, the interconnection words assist of understanding trend flows. The unique words in the specific year strongly related time intervals. Especially, the similarities which have long intervals have low similarities and the similarities which have short intervals have relatively higher values. And, as shown word network distribution using WordNet similarity, the unique word clusters such as IWCs have scattered in same local places. It is easily seen by the word network figures.

5. Conclusion

In this paper, we try to find the research trends using the text mining. We demonstrate the changes of the research trends using the TFIDF results and the comparison of the similarity relationship. In addition, we show graphical representation to more easily understand the distribution of the research trends in the word network distributions.

It can be limited approaches to confirm the research trend if research domains are changed. In this case, we have to re-collect all resources and repeat same progress. In the analysis of research trend, the IWC and the interconnection words are heavily depended on the domain specific problems. Therefore, the analysis of the research trend based on the word similarity still has many obstacles, limitations and challenges.

Extracting common and specific words from terminologies are interesting future works to make the progress precisely and accurately. Extracting the important words using the uniqueness and the interconnection are also significance works to improve the results.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-0001001).

References

[1] Bioinformatics Journal, **(2012)** July 13, http://bioinformatics.oxfordjournals.org.

[2] W. W. Chapman and K. B. Cohen, "Current issues in biomedical text mining and natural language processing", Journal of Biomedical Informatics, vol. 4, no. 5, **(2009)**, pp. 757-759.

[3] A. M. Cohen and W. Hersh, "A survey of current work in biomedical text mining", Briefings in Bioinformatics, vol. 6, no. 1, **(2005)**, pp. 57-71.

[4] W. W. Chapman and K. B. Cohen, "Current issues in biomedical text mining and natural language processing", Journal of Biomedical Informatics, vol. 4, no. 5, **(2009)**, pp. 757-759.

[5] H. Zheng, C. Borchert and Y. Jiang, "A knowledge-driven approach to biomedical document conceptualization", Artificial Intelligent in Medicine, vol. 49, no. 2, **(2010)**, pp. 67-78.

[6] C. Senger, B. A. Gruning, A. Erxleben, K. Doring, H. Patel, S. Flemming, I. Merfort and S. Gunther, "Mining and evaluation of molecular relationships in literature", Bioinformatics, vol. 28, no. 8, **(2012)**, pp. 709-714.

[7] R. Chen, H. Lin and Z. Yang, "Passage retrieval based hidden knowledge discovery from biomedical literature", Expert Systems with Applications, vol. 38, no. 8, **(2011)**, pp. 9958-9964.

[8] J. L. Neto, A. D. Santos, C. A. A. Kaestner and A. A. Freitas, "Document Clustering and Text Summarization", Proceeding of the 4th international Conference Practical Applications of Knowledge Discovery and Data Mining PADD-2000, **(2000)**, pp. 41-55; London.

[9] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries", **(2012)** September 19, http://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf.

[10] R. Saracoglu, K. Tutuncu and N. Allahverdi, "A new approach on search for similar documents with multiple categories using fuzzy clustering", Expert Systems with Applications, vol. 34, no. 4, **(2008)**, pp. 2545-2554.

[11] D. Lin, "An Information-Theoretic Definition of Similarity", In Proceedings of the 15th International Conference on Machine Learning, **(1998)**, pp. 296-304.

[12] J. J. Jiang and D. W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", In Proceedings of International Conference Research on Computational Linguistics **(1997)**, pp. 9-33.

[13] WordNet, **(2012)** July 13, http://wordnet.princeton.edu/.

Authors

KyoJoong Oh

He received a bachelor's degree of computer science in 2011 from Korea Advanced Institute of Science and Technology. He is currently a MS/Ph. D. candidate student in the department of computer science at Korea Advanced Institute of Science and Technology.

Chae-Gyun Lim

He received a bachelor's degree of medical computer science in 2011 from Eulji University. He is currently a research agent in the department of computer science, Korea Advanced Institute of Science and Technology.

Sung Suk Kim

He received a Ph.D. degree of electrical engineering in 2005 from Chung Buk National University. He is currently a research assistant professor in the department of computer science, Korea Advanced Institute of Science and Technology.

Ho-Jin Choi

He is currently an associate professor in the Dept. of Computer Science at KAIST. In 1982 he received a BS in Computer Engineering from Seoul National University, Korea. In 1985 he got an MSc in Computing Software and Systems Design from Newcastle University, UK. And in 1995, he got a PhD in Artificial Intelligence from Imperial College, London, UK. From 1982 to 1989, he worked for DACOM, Korea, and between 1995 and 1996 worked as a post-doctoral researcher at Imperial College. From 1997 to 2002, he served as a faculty member at Korea Aerospace University, Korea. He moved to Information and Communications University (ICU), Korea, from 2002 to 2009. And since 2009 he has been with the Dept. of Computer Science at KAIST. Between 2002 and 2003 he visited Carnegie Mellon University, Pittsburgh, USA, and served as an adjunct professor the Master of Software Engineering (MSE) program. Between 2006 and 2008, he served as the Director of the Institute for IT Gifted Youth at ICU. Since 2010, he has been participating in the Systems Biomedical Informatics National Core Research Center at the Medical School of Seoul National University. Currently, he serves as a member of the board of directors for the Software Engineering Society of Korea, for the Computational

Intelligence Society of Korea, and for Korean Society of Medical Informatics. His current research interests include artificial intelligence, data mining, software engineering, and biomedical informatics.