

Knowledge Extraction from Academic Journals Using Data Mining Techniques

남수현*, 김흥기**

목 차

I. Introduction	V. Application of clustering
II. Previous research on structuring knowledge of academic journals	VI. Link structure analysis
III. Data structures	VII. Knowledge spillover effect
IV. Application of Association	VIII. Conclusion
	References

Key Words : Data Mining, Knowledge Extraction, Link Analysis, Knowledge Spillover

Abstract

최근 우리는 인접학문 간 그리고 학계와 산업계 간의 연구협조가 점차 증가하고 있음을 보아오고 있다. 이러한 현상은 특히 학술저널 간 지식의존성을 촉진하는 계기를 제공하고 있다고 할 수 있다. 본 논문의 목적은 관련저널 간 지식상호 의존성을 규명하고 저널지식의 구조화를 위하여 association, 군집화, 링크분석 등 데이터마이닝 기법을 적용하는 방법론을 제시하는 것이다. 제시된 방법을 통하여 기대되는 점들은 1) 논문의 기본 속성인 키워드, 저자, 그리고 인용데이터를 통합하는 규칙 집합을 통하여 논문지식검색기능의 향상, 2) 키워드를 기반으로 관련 저널 간 그리고 저널내부의 군집분석으로 지식동향 파악, 3) Kleinberg (1999)의 권위와 허브 개념을 인용데이터 분석에 활용하여 기존의 양적 평가 기준인 영향력지수 (impact factor)의 문제점을 보완하며, 4) 특정 논문이나 저널의 지식파급과 관련한 영향력을 산출하는 잠재적 지식파급 지수를 제안하는 것이다.

* 한남대학교 경영정보학과 부교수, namn@hannam.ac.kr, (042) 629-8032

** 한남대학교 국제통상학과 교수, hongkee@hannam.ac.kr, (042) 629-7597

I. Introduction

The process of knowledge creation, diffusion, specialization, generalization, and application depends on the dynamics of the research community network involved: the typology and strength of links among nodes in the network.

Scholastic research papers have been considered as one of the most important sources of powerhouse for R&D activities which in turn differentiate corporate competitiveness from many perspectives such as core technology, product or service conceptualization, and production.

We also notice that the identity of a specific academic area becomes ambiguous since interdisciplinary nature is a popular theme in research community.

Advance in searching technology and database makes it easier for users to retrieve relevant information from diverse scholastic journals. However, the search is usually based on a combination of key words, which doesn't provide much insight for understanding inter-relationship among academic journals or papers. (In this paper there are two meanings of "key word": 1) User-provided key words which are usually given after abstract. Journals require authors to specify small number of key words. Our analysis refers to this type of key words. And 2) Noun-form of words in

the main body of a paper, which are usually employed for text mining. He (1999) followed this definition.

Data mining techniques are useful to find out hidden patterns and trends which are embedded in data, so that the techniques are used extensively in business settings for such purposes as enhanced customer relationships, marketing, customer segmentations, fraud detection etc., (See Berry & Linoff (2004) for more details of applications).

Compared with the emphasis of previous research on studying the evolution and dynamics of science knowledge from ex-ante perspectives, a meaningful guideline for scholars to search appropriate knowledge will be useful for the knowledge creation process.

In this paper we propose a new methodology employing data mining technologies such as clustering, association, and link analysis to identify interrelationship among journals and papers. In specific we use the main attributes of academic paper such as key words, authors and citations. The structure of this paper is as follows:

We provide related previous research in section II. Data structures in consideration are provided in section III. In section IV, based on key word, author, and citation data, we apply the association technique to making knowledge searching mechanism more effective by generating rules. We show

that how clustering techniques are applied to dealing with key word in section V, followed by link analysis for identifying influential papers using citation data in section VI. In section VII knowledge spillover effect of a paper and journal is proposed to measure the potential of knowledge diffusion and the conclusion in VIII is reached.

II. Previous research on structuring knowledge of academic journals

Key components of academic paper consist of title, authors, key words, main body and citations. Of course the main body of papers is the essence. However considering the evolutionary and cooperative nature of research, there should be some methods to identify the relationship among journals and papers. To identify the underlying relationship, a set of attributes such as authors, key words, and citations are frequently used since they contain condensed information representing the knowledge.

Lately, stimulated by bio-informatics, text mining techniques are gradually employed for a detailed and extensive analysis of the main body (Sullivan, 2001).

This method might be effective for both detecting important links between related research and to discovering knowledge itself from a huge set of papers.

The "impact factor" (Garfield, 1999) has been widely used to differentiate journals. The impact factor for a journal is obtained in the following way: Let $C_{j,t}$ be the total number of citations which refer to journal j during t and $C_{j,t-b}$ be the number of citations to journal j during the previous b years, or between the years of $t-b$ and $t-1$. Notice that $C_{j,t} \geq C_{j,t-b}$. Also define $p_{j,t}$ to be the total number of papers published in journal j during $t-b$ and $t-1$. Then, the impact factor of time window of b years for journal j and year t , $IF_{j,t}^b$, is calculated as

$$IF_{j,t}^b = C_{j,t} / p_{j,t-b} \quad (1)$$

Two-year time window ($b=2$) is generally accepted as a good indicator to measure the influence of a journal. However if a journal deals with subjects which don't decay fast, a bigger b may be used. Notice the number of citations during the time window tends to increase as the frequency of publications of the journal is high and the number of papers in a volume of that journal is bigger. But this effect is offset by the denominator. To measure an appropriate b the cited half life can be used. The half life is obtained by finding b such that 50% of $C_{j,t}$ are included in the previous b years.

Impact factor is to measure how

influential a journal is. However, the influence is based mostly on the magnitude of counting not on the quality of citation. The concept of impact factor can be applied to authors and individual papers, but the citation records for these are much less than for journals.

Since impact factor is largely based on the simple counting of citations, the relevance of the measure is often questioned (Amin and Mabe, 2000).

Kleinberg (1999) observed that the quantitative concept of impact factor might mislead the importance of a journal or a paper, by differentiating the importance into three category: authority, hub, and popularity. His algorithm was originally adopted by Internet search community such as Google, but it can be easily extended to scholastic community too.

Author attribute was used to measure research collaboration (Hicks et al., 1996). They compared the degree of research collaboration of firms in Japan and Europe in time series, examining co-authors of research papers. They found firms in Europe were more active in the collaboration than in those in Japan. Moreover, the gap between the two regions were getting bigger, indicating different technological opportunities available to firms in the two regions.

Jaffe et al. (1986) shows how to use R&D spillover using patents which require a

rigorous citation by law (Jaffe and Trajtenberg, 1996). He explained the amount of knowledge creation in terms of R&D efforts and potential spillover pool at firm level. Knowledge creation was measured by the number of patents; R&D efforts by R&D expenditures; and spillover pool by the proximity of a firm's knowledge creation weighted by R&D efforts by other firms. We import the idea of Jaffe(1986) to measure the potential knowledge spillover of a paper, an author or a journal using citation and impact factor information.

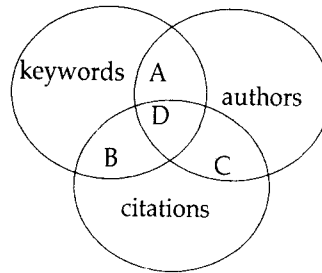
Advances in computer processing capability make co-word analysis feasible and widely used to extract knowledge from the text and to find linkage among arbitrary papers (He, 1999). The analysis does not depend on any priori definition like user-provided key words or citations. Assuming that the words in the main text are the "actors", the analysis can be used to provide a diverse set of quantitative associations among words. The central theme of co-word analysis is to find the inclusion and proximity indexes which are used to determine the major and minor research areas. From the indexes, inclusion and proximity maps are created to view the linkage among research areas.

However, as we see in the previous research, the three important dimensions of author, key word, and citation are not utilized in integrative ways. Therefore in this paper we present ways of extracting

knowledge from integral point of view as in Figure 1.

citation, with cardinality of α , k , and c , respectively. From this, we create a vector

⟨Figure 1: Interaction of three attributes of a paper⟩



In section IV we extend the co-word analysis by accommodating intersecting area such as A in ⟨Figure 1⟩ and generating rules so that a richer information can be extracted using association techniques.

III. Data structures

To explain our methodology we use a vector which consists of 3 elements: author supplied key words, citations, and authors.

Consider an imaginary set of j journals whose subject are categorized as similar. Then we may assume that there exists some similarity in key words, authors, and citations among journals in the group. We consider a certain period of time window. In that time frame we prepare three sets such as A for author, K for key word, and C for

of size which is the sum of α , k , and c . Let T_W be a paper p in journal j . Then each element of T_W is either 0 or 1, depending on whether the paper has the corresponding attribute element. All the subsequent analyses depend on T_W .

To show the data structures visually, consider the following fictitious data for papers on journals, key words, authors, and citations as in ⟨Table 1⟩. Since a paper can only cite papers published before, the citation is time dependent. Also note that a paper can cite many papers and references other than those in the collection. To reduce the size of the vector space, we can consider the citations to those papers in the collection; in the following example, p_a and p_b are outside the collection. The vector space in the collection consists of $j_1 - j_2$, $p_1 - p_5$, $a_1 - a_4$, $k_1 - k_3$, and the citation space of $p_1 - p_5$.

<Table 1: Example of Source Data>

journal	paper	author	key word	citation
j1	p1	a2	k3	pa
j1	p2	a1 a3	k1 k2	p1
j1	p3	a4	k1	p1 p2
j2	p4	a3	k2 k3	pb p1 p2
j2	p5	a1	k2	p1

Then the first, third, and fourth rows of <Table 1> can be compiled as vectors, $T_{k,i}$, in <Table 2> based on whether a paper has the corresponding attribute. If the attribute exists, assign 1, and 0, if it doesn't.

together. However, it does not convey any cause and effect relation. If there is a high tendency that beer triggers the purchase of salted peanuts, then the relation can be thought of as a rule. A set of rules can be effectively used for promotion, shelf

<Table 2: Compiled Data Structure>

ID	set A				set K			set C				
	a1	a2	a3	a4	k1	k2	k3	p1	p2	p3	p4	p5
j1p1	0	1	0	0	0	0	1	0	0	0	0	0
j1p3	0	0	0	1	1	0	0	1	1	0	0	0
j2p4	1	0	0	0	0	1	1	1	1	0	0	0

IV. Application of Association

In business world, association techniques are widely used in customer relationship management. Association means that some events occur simultaneously, as in the purchase of beer and salted peanuts

arrangement and layout, and recommendation to customers, especially in Internet shopping environment.

Several methods for analyzing the degree of association of key words are proposed in He (1999). Two related major indexes are inclusion index (I_j) and proximity index (PR_j) defined in the following way:

$$I_j = n(k_i, k_j) / \min n(k_i), n(k_j) \quad (2)$$

where

$n(k_i)$ = the number of papers containing the key word k_i ;

$n(k_j)$ = the number of papers containing the key word k_j ;

$n(k_i, k_j)$ = the number of papers containing key words of both k_i and k_j .

$\min(a, b)$ = the minimum of two quantities, a and b .

Also the proximity index is defined as,

$$PR_{ij} = n(k_i, k_j) \cdot N / (n(k_i)n(k_j)), \quad (3)$$

where

N = the total number of papers in the set.

Now let's introduce the notions of support, confidence, and lift (See Berry and Linoff, 2004 for more detailed description). Note that (2) and (3) just measure the strength of types of combination without any directional consideration.

Suppose now that a rule, if k_i , then k_j , stating that if a paper has a key word k_i , then that paper also contains key word k_j with certain probability. Call the probability of co-occurring k_i and k_j the support of the rule or $S(k_i, k_j)$ which can be expressed as

$$S(k_i, k_j) = n(k_i, k_j)/N, \quad (4)$$

Note that $S(k_i, k_j)$ applies to a rule, if k_i ,

then k_j , in the symmetric way. But to measure the effectiveness of an rule by taking into account of the direction, we need another probability. We call it "confidence". For the rule, if k_i , then k_j , define the confidence,

$$C(k_i \rightarrow k_j) = S(k_i, k_j)/P(k_i), \quad (5)$$

where $P(k_i) = n(k_i)/N$. It is easy to see that (5) is a probability conditioned on the set of papers containing key word k_i .

Now compare equations (2) and (5). Equation (5) undershoots due to the denominator. However the main difference is that equation (2) does not convey any directional information, while equation (5) does. Moreover, He (1999) does only consider the inclusion index between two key words, but we can extend (5) to a general case of a rule such that if $k_1 \& k_2 \& \dots \& k_n$, then k_z .

To determine how much better a rule performs we need another measure, "lift", which is defined as (see Berry and Linoff, 2004, pp 310):

$$L(k_i \rightarrow k_j) = C(k_i \rightarrow k_j)/P(k_j), \quad (6)$$

Equation (6) measures the ratio of the rule's confidence, $C(k_i \rightarrow k_j)$, to the random probability of getting key word k_j , $P(k_j)$.

$L(k_i \rightarrow k_j) > 1$ implies that the rule is significant. But if $L(k_i \rightarrow k_j) \leq 1$, then the

rule is not meaningful.

The equations (3) and (6) are the exactly the same. But equation (3) not only does not contain directional information but also cannot be extended to many key words situations.

Based on the equations (5) and (6), we can provide scholars with more flexible ways of searching mechanism by way of rule generation. The rules can include elements from the sets of A , K , and C . For example, if there is a rule base stating that

Rule 1 : If k_i , then k_j .

Rule 2 : If $k_i \& k_j$, then a_i .

and the user searches related information k_i , then the search engine can search the rule base to provide both key word k_j and the author a_i .

To accommodate all the possible combinations of rules, we need to overcome the exponential growth of problem space derived from the vector size w . The complexity of this problem is non-polynomial. There are several ways of handling this problem: 1) For a rule to be useful, a threshold for the support can be predefined. That is, if $S(k_i)$ is less than the threshold, any further combination including k_i can be eliminated from consideration, which significantly reduces the size of problem space. 2) To handle the rule generation process efficiently in terms

of memory space and time, we may use the apriori algorithm (see Han & Kamber, 2001).

3) If there are many key words, the key words can be aggregated into more general key words so that the vector size is reduced.

V. Application of Clustering

In section III we assumed that the journals in consideration are similar in terms of research area. However, there might be intrinsic differences among journals due to the differences in editorial policy and research community among journals. If that is the case, the extraction of major research area of each journal can be useful for researchers. In this section we propose techniques both to test whether the journals in the set deal with the same research area or not and to provide clusters of research subjects for each journal.

User supplied key words convey condensed meanings of the paper. Analysis of the key words can give us such information as research area, research methodology, and so on. Given a set of j journals, it is interesting question to ask whether those journals are different with respect to key words. If there is no

significant difference, editorial board of each journal must try to figure out its identity. Moreover, if a governmental level organization oversees the journals, funding to those journal might not be effectively utilized.

Testing the hypothesis

Hypothesis: There is no significant differences of research subject among the journals in consideration.

To test the hypothesis we need to rearrange the vectors $T_{p,j}$ so that only the key word portion remains. Note that the element of the vector of $T_{p,j}$ is 1, if paper p in journal j contains key word k_i and 0, otherwise.

Since we don't have any priori knowledge on key words, an agglomeration clustering technique (see Han and Kamber, 2001 for details) is better to identify the appropriate number of clusters for key words. General idea of finding a good cutting point for agglomeration clustering is to build up clusters from scratch so that the homogeneity of a cluster is high and the between clusters difference is high. One way of choosing the number of clusters is the clustering coefficient in SPSS. Choi & Lee (2002) recommends the number which makes the slope of the clustering coefficient maximum.

From this procedure assume that we have come up with g clusters. To test the null hypothesis we build a contingent table with rows and columns representing g clusters and j journals, respectively. For each cell we assign two numbers: one is the real number of papers for each cluster and journal; the other is the expected number of occurrence. From the table filled up with cells of two numbers, a χ^2 test with the degree of freedom, $(g-1)(j-1)$, can be performed.

If the null hypothesis is rejected, we can look into detail of each journal to specify the research area. For this analysis we may use a top down clustering algorithm such as K-means. We can choose an arbitrary number of K.

VI. Link structure analysis

1. Problems of assessing importance of journals and papers

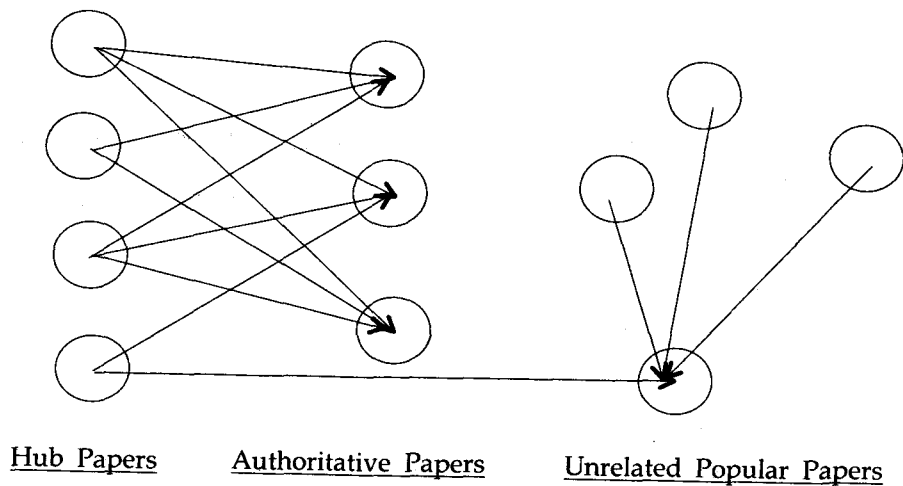
As we saw the impact factor for a journal or a paper is based on the counting of citations to that journal or paper. The number of hits at Internet sites had been considered an important factor to evaluate the sites, making it a baseline for assessing

advertisement fees. However, the number itself does not reflect the real value of the sites (Hoffman and Novak, 2000). Also all hits are not equally important: some hits are more important than others in terms of purchasing propensity. This phenomena can be applied to scholastic papers. In this sense, the impact factor is not a good indicator to measure the paper's influence. Kleinberg (1999) proposed a new way of determining the importance of a webpage. The main observation by Kleinberg is that the importance of paper j depends on both

the amount of citation received and the importance of those citations. Note the recursive relationship of importance between paper j and the citing papers.

The key idea of the Kleinberg's algorithm is to differentiate papers into three classes: the first group is represented as authoritative; the second as hub; and the third as popular (Berry and Linoff, 2004) as in <Figure 2>.

<Figure 2: Three classes of papers,
Revised from Berry & Linoff, 2004, p 335>



2. Application of link structure analysis on scholastic journals

We provide a brief procedure of the derivation of these sets of papers (see Chakrabarti et al., 1999 for detailed description of the HITS algorithm which is implemented in statistical packages such as SPSS).

Consider the portion of citation data from the vector $T_{p,j}$. That portion of data refers to the existence of citation to other papers. These other papers may be included in the set J or not. For efficient processing we can limit the size of citation vector by considering those papers in J only (See section III). Thus when a paper in J cites another paper not in J , that cited paper is not included in constructing the adjacency matrix. Suppose the total number of papers in J during the time window of consideration is n . Define the adjacency matrix as C , a $n \times n$, square matrix. Also define $C_{ij} = 1$, if paper i cites paper j , and 0, otherwise. Then the relations between the vectors of authoritative (x) and hub (y) can be written as:

$$x = C^T y. \quad (7)$$

$$y = Cx \quad (8)$$

Note that C^T is the transpose of C . Since C_{ij} represents the existence of citation from paper i to paper j , equation (8) captures the outgoing (hub) force. In similar way

Equation (7) can be interpreted as incoming (authority) strength. Also note that Equations (7) and (8) are recursively related. So substituting each other we can rewrite these two variable equations into a single variable ones as follows: $x = (C^T C)x$, and $y = (C C^T)y$. Solutions for both x and y can be obtained numerically by these equations.

We can apply Kleinberg's algorithm either on the citation data obtained from across the set J or on the citation data created from the clusters of the set J .

VII. Knowledge spillover effect

1. Background

Academic papers are considered as the important sources of codified knowledge. Since the knowledge is made public, it has many aspects of a public goods. Moreover, the existence of related knowledge published by other scholars makes a given researcher put much less efforts than he creates the knowledge from the scratch. In general, if research subjects and ideas of paper are promising, the papers induce many other subsequent papers, which we call the phenomena of a knowledge

spillover. Impact factor and authoritativeness can be proxies to measure the degree of knowledge spillover contributed by a paper. The main characteristic of the set of papers of both inducing and induced spillover is that their proximity is very close. In this section we propose a measure of proximity so that we can measure the importance of a paper from the knowledge spillover perspective. Jaffe (1986) derived a model to estimate R&D productivity at firm level. His conjectured that if any two firms have a close proximity of patent composition across the 49 patent categories, the potential of R&D spillover will be high, which in turn stimulate the given firm's R&D productivity. He called the potential as "potential R&D spillover pool", whose concept will be used for our measure of knowledge spillover effect.

2. Measuring knowledge spillover with citation

We measure knowledge spillover of a paper or a journal using citation information. To do it, we deal with the citation matrix, C introduced in section VI. Select any two rows, C_i and C_j from C . Then the similarity of the two rows (s_{ij}) can be obtained by the cosine angle of the two vectors as follows:

$$s_{ij} = \cos \theta_{ij} = \frac{C_i \cdot C_j}{\|C_i\| \|C_j\|}, \text{ where } (9)$$

the numerator denotes the inner product of two vectors and it is normalized by the product of the norms. Note that $0 \leq \cos \theta_{ij} \leq 1$; when the two citation vectors have the exactly same components, the cosine value is 1, and 0, otherwise. We believe the angular separation measure (see Sullivan, 2001, for measuring the separation of vectors with text data) of similarity as in Equation (9) is better than the Euclidean distance since it is more computationally efficient due to the sparsity of the matrix and the composition of elements of the vector is more important than the length of the vector.

Another index affecting the magnitude of spillover is the degree of importance of papers. We can use the authoritativeness defined in section VI or impact factor in section II. Denote the importance of paper as w_j . Based on the similarity index between any two papers and the importance index, the potential knowledge spillover pool into the paper i (KSP_i) is given as:

$$KSP_i = \sum s_{ij} w_j. \quad (10)$$

The index, KSP_i , is the product of the similarity between the two papers with respect to citation and it is weighted by the

importance of the neighboring paper. So the index measures the gravity into the given paper from the neighboring environment. If it is high, it is conjectured that the externality of that paper is high.

This concept can be extended to the potential knowledge spillover induced by authors and journals by aggregating the citations of papers according to authors or journals.

VIII. Conclusion

We extended the previous related research on knowledge discovery from academic journals by employing data mining techniques such as association, clustering, and link analysis. In specific, we extended the domain of simple key word searching to

that of rule based searching which integrates diverse dimensions such as key words, authors, and citation information. We also augmented the counting based impact factor with authoritativeness of citation.

Utilization of scholastic papers are important for the governmental research policy setting toward R&D and promoting cooperation between academic institutes and industries. In this sense the knowledge spillover effect should be considered as important factor for assessing papers, journals, and authors. To do that we introduced the concept of potential knowledge spillover pool using citation information.

Considering the recent governmental initiative for Digital Library which requires to provide higher utility to users through efficient knowledge creating efforts as well as enhanced and customized searching capability, we hope this research brings attention to further research.

References

1. Amin, M and M. Mabe, "Impact Factors: Use and Abuse", *Perspectives in Publishing*, October, 2000, (www.elsevier.com/framework_editors/pdfs/Perspectives1.pdf).
2. Chakrabarti, S, B. Dom, S. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, J. Kleinberg, "Mining Web's Link Structure", *Computer*, August, 1999, pp 60-67.
3. Choi, B. and H. Lee, "Knowledge Management Strategy and its Link to Knowledge Creation Process", *Expert Systems with Applications*, Vol 23, pp 173-187, 2002.
4. Garfield, E., "Journal impact factor: a brief review", *Canadian Medical Association Journal*, Vol 161, No 8, pp 979-980, 1999.
5. Han, J. and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.
6. He, Q., "Knowledge Discovery Through Co-Word Analysis", *Library Trends*, Vol 48, No 1, pp 133-159, 1999.
7. Hicks, D, P. Isard and B. Martin, "A morphology of Japanese and European corporate research networks", *Research Policy*, Vol 25 No 3, pp 359-378, 1996.
8. Hoffman, D. and T. Novak, "How to Acquire Customers on the Web", *Harvard Business Review*, Vol 78, No 3, pp 178-188, 2000.
9. Jaffe, A., "Technological Opportunities and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value", *The American Economic Review*, Vol 76, No 5, pp 984-1001, 1986.
10. Jaffe, A. and M. Trajtenberg, "Flows of knowledge from universities and federal laboratories: Modeling the flow of patent citations over time and across institutional and geographical boundaries", *Proceedings of National Science*, Vol 93, pp 12671-12677, 1996.
11. Kleinberg, J, "Authoritative Sources in a Hyperlinked Environment", *Journal of the ACM*, Vol 46, No 5, pp 604-632, 1999.
12. Sullivan, D., *Document Warehousing and Text Mining*, Wiley, 2001.