The Impact of the Latent Semantic Analysis on Science and Technology: A Bibliometric Analysis

Hamid Darvish
PhD. student at Information Management, Hacettepe University
Ankara, Turkey
Instructor at Computer Engineering Department, Cankaya University
Ankara, Turkey

Abstract:

Latent Semantic Analysis (LSA) has been in use in several different fields of science. Several modeling techniques including Boolean, set-theoretic, vector space, and probabilistic models studied. In this paper, we first describe the concept of "LSA" and then present the preliminary results of an exploratory study. We carried out a small-scale bibliometric analysis to find out the impact of LSA on various scientific and technological fields. We downloaded bibliographic records with "Latent Semantic Analysis" in their titles from Thomson's Science Citation Index Expanded and used Bibexcel and Pajek to perform several bibliometric and network analyses such as co-citation, co-authorship and co-word. It appears that *LSA* has had an impact on a wide variety of scientific disciplines from discourse analysis to cognitive science to machine learning.

Keywords: Bibliometric Analysis, Latent Semantic Analysis, Social network, Latent Semantic Indexing

Introduction:

Susan T. Dumais and Thomas K. Landauer, et al [1], patented LSA in 1988. Since then, *LSA* has been appearing in numerous journals, for example, Information science & Library science, Medical informatics, Discourse Analysis, Computer science & Information systems and so on.

LSA is fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passage of discourse. It is not a traditional natural language processing or artificial intelligent program; it uses no humanly constructed dictionary, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies. One of the applications that use the LSA concepts is information retrieval. Latent Semantic Indexing (alias for LSA) is an information retrieval method that matches the term query with the related documents.

As J.R., Anderson states that, "the analogy between information retrieval and human semantic memory processes" [2]. Hence, LSA maps the analogy, literary; LSA matches a word in the mind of a user with a text that has the same meaning in the system (Database, Internet). Search engines (Yahoo, Google), have included the LSA in their algorithms as well. Applications of LSA range from Information filtering, Cross Language Information Retrieval, to score TOEFL (Test of English as a Foreign Language) tests.

Advantages of latent Semantic Indexing

Information Retrieval suffers from synonymy and polysemy [3]. Former one means that that an object seen in many ways, an example, is words automobiles and cars whereas the polysemy is having different words with the same meaning. For example, words such as apple, as fruits, or apple as a computer. LSA overcomes this problem by using Singular value decomposition (SVD). Latent Semantic Indexing is a well-known technique in Information Retrieval, especially in dealing with polysemy and synonymy. LSI uses SVD process to decompose the original term-document matrix into a lower dimension triplet. The triplet (the resulted matrices) is the approximation to original matrix and can capture the latent semantic relation between terms [4].

According to the project, "*Document Ontology Extractor (DOE)",* done by the Research Team, Govind R Maddi, Jun Zhao, Chakravarthi S. Velvadapu form University of Maryland Baltimore County, Bowie State University with the sponsoredship of Department Of Defense, USA, Singular Value Decomposition decomposes a given matrix into three components - U, S and V[5].

m x n term-document matrix **A**, of rank **r**, expressed as the product:

- $\mathbf{A = U * S * V^T}$ (1)
- **U** is m x r term matrix
- **S** is r x r diagonal matrix
- **V** is r x n document matrix
- Diagonal of S contains singular values of A in the descending order.
- **A** is formed from LSI as follows:
- $\mathbf{A = U^S * S^S * V^{sT}}$ (2)
- $\mathbf{U^S}$ - derived from **U** removing all but the **s** columns
- $\mathbf{S^S}$ - derived from **S** removing all but the largest **s** singular values
- $\mathbf{V^{sT}}$ - derived from $\mathbf{V^T}$ removing all but the **s** corresponding rows

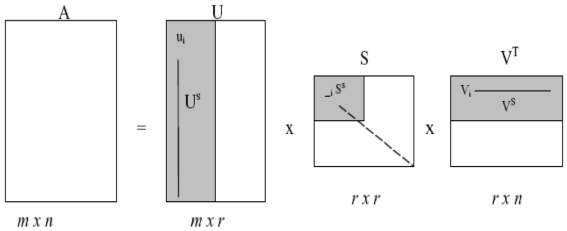

Figure 1 Singular Value Decomposition Demonstration

Author(s)	Observed Citation
1	2896
2	417
3	134
4	65
5	35
6	19
7	10
8	6
9	7
10	6
11	3
12	4
13	3
14	1
15	3
17	2
18	1
19	1
21	1
22	2
28	1
31	1
35	1
37	1
39	2
55	1
72	1
105	1
178	1

Table 1 Scientific Productivity for authors (n=237)

Briefly, the basic idea is that *LSA* treats a passage as a linear equation; its meaning well approximated as the sum of the meaning of its words using a matrix [6].

$$\mathbf{m\ (passage) = m(word_1) + m(word_2) + m(word_n)}$$
$$\mathbf{m\ (psg_i) = m(wd_{i1}) + m(wd_{i2}) + \ldots + m(wd_{in}) \quad (3)}$$

Disadvantages of LSA:

In nutshell, the term query and document presented as vector space, as the document gets bigger the storage and efficiency becomes a problem. However, some argue otherwise.

Methodology:

Since "a picture is worth a thousand words", in this paper, we have used Bibexcel to refine the raw data from Thomson's Science Citation Index Expanded and Pajek constantly to draw the network diagram graphically. The network analysis reveals the relationship among authors whom disseminate knowledge through their cooperation. Social network were introduced by Moreno [7]. He states that actors play important role in the sociogram (network diagram). Sociogram consists of nodes and edges. Each node (actors) and edge (linkage among actors) represents a betweenness among authors. Actors joined with lines representing ties, as in a social network [8]. The key word LSA in the selected sample has been studied. By utilizing bibliometric methods, such as co-authorship, co-citation among journals, indicates the significance of LSA in science and technology. Moreover, by applying the Zipf"s law we measure the frequency of the word *LSA* within sample journals. In addition, according to facts from Thomson's Web of Science (WoS) imply that on the sample consisting of 237 articles based on the key word *LSA* has an h-index factor of 19 and median of citation per item 9.22.
Bibliometrics means "the application of statistical and mathematical methods to books and other media of communication" [9]. The sample consists of 237 articles from Thomson's Web of Science (WoS) that contained the word "*LSA*" in their titles or as a keyword in their abstracts. The time span is 1986-2007, collected from Thomson's Web of Science (WoS). Furthermore, co-word analysis of the keywords *LSA* with other keywords used in the sample articles, using Zipf's law, reveals certain word patterns with LSA key word. In addition, applying Lotka's law reveals that scientific productivity for the scientists whom have used *LSA* directly or indirectly in their research.

Findings:

The frequency of LSA key word that appeared in the samples top ten journals, it is an indicator of usage of the term in the sample article. An example, the LSA has appeared 146 times in the journal of the "Discourse Process" whereas 37 times has appeared in journal of the "MACH LEARN".

Table 2. Co-occurrences of LSA in Journals Article (n=237)

Journal Name	Number of Articles
DISCOURSE PROCESS	**146**
PSYCHOL REV	**131**
J AM SOC INFORM SCI	**105**
BEHAV RES METH INS C	**44**
COGNITIVE SCI	**40**
MACH LEARN	**37**
J MEM LANG	**37**
J EXP PSYCHOL LEARN	**34**
COGNITION INSTRUCT	**33**

Applying some informatics "law" such as Zipf's laws for the frequency of words in below table, the rank-size yields: The "size" is the key word frequency, the rank (1 for the highest rank, 2 for the second and so on...)

Key Words	Freq	Rank	Ln rank	Ln Fre
LSA	117	1	0	4.76
KNOWLEDGE	41	2	0.69	3.71
REPRESENTATION	17	4	1.09	2.83
COMPREHENSION	17	6	1.38	2.83
MODEL	16	7	1.60	2.77
MEMORY	14	8	1.79	2.63
TEXT	14	9	1.94	2.63
ACQUISITION	12	1	2.07	2.48
RECOGNITION	9	1	2.19	2.19
INFORMATION-RETRIEVAL	8	12	2.30	2.07

Table 3. LSA word Frequency compared with other key words

The graph below shows ln (size) on the vertical axis and ln (rank) on the horizontal axis. The Alpha is close to 1 so, implies Zipf's law.

$$\ln (\textit{size}) = \text{constant} - \alpha \ln (\textit{rank}) \quad (4)$$

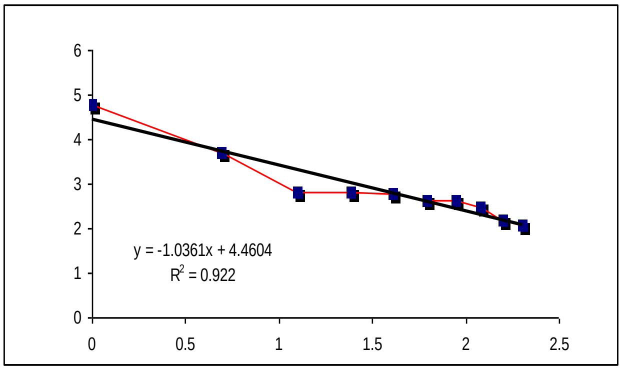

Figure 2 Zipf's law demonstration Logarithmic Result

Another indicator that shows the significance of *LSA* is to look into the co-authorship among authors whom cite each other in their research process. The network diagram of the first authors who co-authored at least three articles on *LSA* published in various journals is given in figure 3. The sociogram depict that the status, centrality of Landauer cluster and its relation with his neiborhood. Nonetheless, since the actors are the same type, it is one-mode type sociogram. Network diagram reveals a certain pattern among researchers such that scholarly papers published by Landauer's cluster and Graesser's cluster are central foci of the sociogram, whereas, Magliano JP, Millis KK and Wiemer-Hasting K appear as a bridge between two main clusters. We might include that there is transitivity between two clusters.

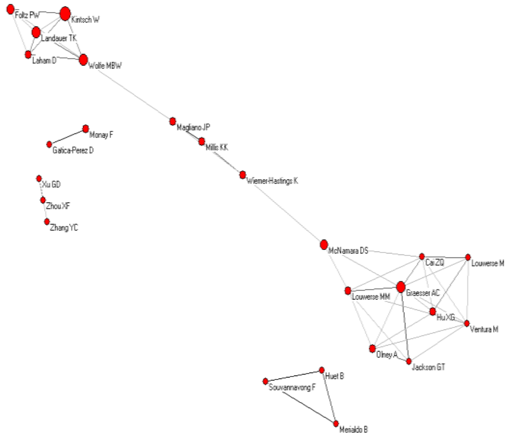

Figure 3 Co-authorship network on LSA

There are three isolated clusters and three connected clusters. Landauer's cluster connected to Millis cluster through Magliano. That cluster connected to Graesser cluster through McNamara. Each author studies on a specific field. For example, the Kitsch whose work is essential to LSA, which appears in Journal of Discourse analysis, has the highest cited author in his cluster; Landauer follows next. Moreover, each cluster shows the collaboration among authors closely related to the discipline, which they have done research. On the other hand, isolated clusters represent another field of research, for example, Monay F, and Garcia-Perez D research on multimedia application utilizing LSA. Nevertheless, network diagram depicts the scientific collaboration among authors, from connected to isolated clusters.

Co-citation like the previous one is another example of social network analysis. The sociogram is scoicentric; three authors whose status play important role in the scoicentric network. Co-citation network determines the relationship between authors (1st author). At the center of the graph we see the most congested area consist of Deerwester, Scott; Dumais, Susan T, Graesser AC, Foltz, Barry MW (the inventor of LSA) whom are active in Discourse Analysis research. Since the mentioned authors patented the LSA their role and status has had impact on other researchers as well. As we spread out form center, we observe new names such as Gerald Salton who is consider the father of information retrieval. As we go further, to the edge of the graph, we see name such Baeza-Yate R. who also well known in field of information retrieval. Therefore, the network analysis shows the utilization of the LSA from Discourse Analysis to Information Retrieval.

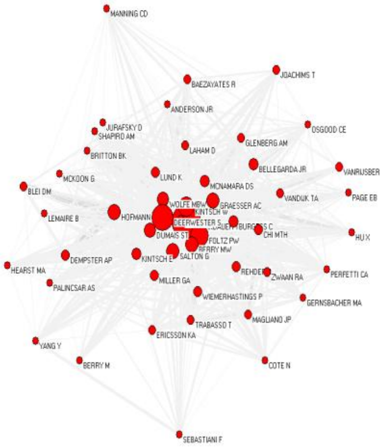

Figure 4 First author Co-citation analysis

Nonetheless, we can conclude that many researchers have cited simply the original work even though there is no direct closeness between original inventors and new researchers.

Unlike the pervious sociograms, in the next sociogram, the “actors” are the key term “LSA” and its co-occurrence in the journal’s articles instead of the authors. We still observe the co-occurrence of LSA in sample articles (the highest ones). “Discourse Process” leads in the center; “J AM Soc Inform Sci” and “Psycho Rev” follows it. Nevertheless, “cognitive Sci”, “Commun ACM”, “Mach Learn” appear in the network analysis as we reach the arc of the circle. In addition, Line between nodes as well as their thickness indicates the number of co-occurrences.

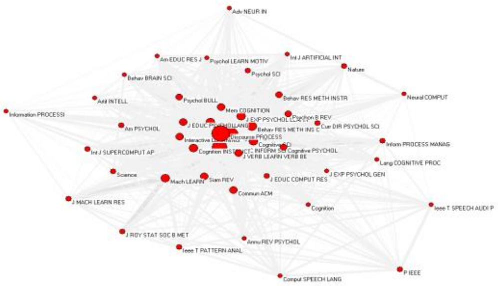

Figure 5 Co-occurrence of the key word LSA

The co-word analysis based on co-occurrences of the key word that appears in bibliographic database form Thomson’s Web of Science (WoS) determines the topicality of the sample articles. In our case, we utilize the DE field of sample articles. There are at least two clusters in the graph. One is which has LSA at its center. Surrounded by, singular value decomposition, information retrieval, statistical language modeling. Sociogram made of two parts. One is an egocentric sociogram which has LSA in the center, co-occurring with words such as n-gram, singular value decomposition, information retrieval. Other one is a scoicentric consists of image retrieval, object recognition, which is a strong indicator of usage of the LSA in Multimedia technology.

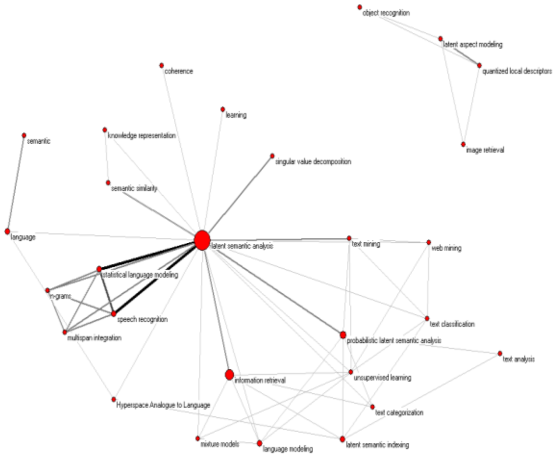

Figure 6 co-word network analysis

Co-citation and co-authorship, co-word network diagram summarize the fact that exists among research communities as follows:

- Knowledge sharing through network of friends
- Portraying the data in term of matrixes
- A graph or sociogram are co-occurring
- Mathematical operation on Matrix reveal the relationship among the actors in the graph

Conclusion:

Applying bibliometrics methods on bibliographic data records of 237 from Thomson’s Web of Science (WoS) revealed that not only LSA has had impact in Discourse Analysis, but also, it has influenced other scientific communities as well. For example, Information Retrieval, Image Retrieval, Machine Learning (name just a few). Another significance of the *LSA* is that, with the advent of search engines (Yahoo, Google); commercial companies are trying to use *LSA* technology in their search engine algorithms. We foresee the scientific citing for *LSA* field in coming future promising in search engines improvement, image and multimedia retrieval as well.

References:

1. Anderson, J. R. (1990). The Adaptive Character of Thought. Hillsdale, NJ: Erlbaum.

2. 2. Landauer, T. K. , Foltz, P. W. , & Laham, D. (1998). Introduction to Latent Semantic Analysis.
3. Discourse Processes, 25, 259-285

4. Deerwester, Scott; Dumais, Susan T.; Furnas, George W.; Landauer, Thomas K. and Harshman, Richard. Indexing by Latent Semantic Indexing. Journal of the American Society for Information Science. 41(6), p. 321-407, 1990.

5. Chung-Hong Lee, Hsin-Chang Yang, Sheng-Min Ma, A *Novel Multilingual Text Categorization System using Latent Semantic Indexing*, **Seen in 24.12.2007** http://doi.ieeecomputersociety.org/10.1109/ICICIC.2006.214

6. Govind R Maddi, Jun Zhao Chakravarthi S Velvadapu“*Document Ontology Extractor (DOE)”,* University of Maryland, Baltimore County, Bowie State University with the sponsoredship of Department Of Defense retrieved in 24.12.2007 at <http:// www.csee.umbc.edu/cadip/2001Symposium/slide1.ppt>

7. "Introduction to latent Semantic Analysis." Simone Denis, Tom Landauer, Walter Kintsch, Jose Quesada: Cogsci 2003 LSA toturial web site. Colorado University. <http:// lsa.colorado.edu>

8. Pritchard, A. (1969). Statistical bibliography or bibliometrics? Journal of Documentation, 25(4), 348-349.

9. Moreno, J. L. (1934). Who Shall Survive? Washington, DC: Nervous and Mental Disease Publishing Company. Pritchard, A. (1969). Statistical bibliography or bibliometrics? Journal Of Documentation, 25(4), 348-349.
10. “Intro to Social Network” John Canny: HCC class lecture 21. UC Berkeley. <http:// **www.cs.berkeley.edu/~jfc/hcc/courseSP05/lecs/lec21/jfc21.ppt**>