Tapping into Scientific Knowledge Flows via Semantic Links

Saeed-Ul Hassan[1] and Peter Haddawy[2]

[1] *saeed-ul-hassan@itu.edu.pk*
Information Technology University, 346-B, Ferozepur Road, Lahore (Pakistan)

[2] *peter.had@mahidol.ac.th*
Faculty of ICT, Mahidol University, 999 Phuttamonthon 4 Rd, Salaya, Nakhonpathom 73170 (Thailand)

Abstract
We present a new technique to semantically analyze knowledge flows between countries by using bibliometric data. Using a new approach to keyword-based clustering, the technique identifies the main topics of the research output of a country, as well as the main topics of the citing research of other countries. In this way it provides insight into how research produced by one country is used by others. We present a case study to illustrate the use of our proposed technique in the subject area of Renewable Energy during 2005-2010 using data from the Scopus database. We compare the Japanese and Chinese papers that cite the scientific literature produced by researchers from the United States in order to show the difference in the use of same knowledge. While the Japanese researchers focus on research areas such as efficient use of Photovoltaics and Superconductors, Chinese researchers focus in areas related to Power Systems, Power Management and Hydrogen Production. Such analyses may be helpful in establishing more effective multi-national research collaboration.

Conference Topics
Methods and techniques; Country-level studies

Introduction
The research collaboration facilitated by the Internet and the greatly increased global mobility of researchers have resulted in a new highly dynamic global marketplace for ideas. The possession of knowledge, the value of which depreciates at an increasingly rapid rate, is no longer as valuable as the ability to participate in the knowledge flows associated with these marketplaces. As observed by Hagel et al. (2009) in the context of business competitiveness, "Knowledge flows – which occur in any social, fluid environment where learning and collaboration can take place – are quickly becoming one of the most crucial sources of value creation". Similarly in Science, understanding a research landscape increasingly requires understanding the dynamics of the relevant knowledge flows.

International scientific leadership and influence are commonly viewed as important measures of a country's scientific intellectual strength. This has traditionally been measured in terms of international scientific collaboration and the ability of a country to attract strong researchers and graduate students from abroad. But a further, more direct measure is the extent to which results generated by a country's researchers are influencing and being utilized by researchers abroad, particularly researchers who are not yet directly collaborating with that country's researchers.

In this paper we present a new technique to measure and semantically analyze knowledge flows between countries by using publication and citation data. We select a set of papers authored by the researches of a given source country. Further, we identify the papers cited by the papers only authored by researchers from outside the source country. We cluster these internationally cited papers to identify the main topics. Then, we procure the sets of papers (authored by researchers outside the given country) citing each of the topic clusters. Finally, we in turn cluster each set of citing papers to again identify main topics in order to identify how the knowledge from the topics in the cited papers is being used.

Related Work

In bibliometrics there have been efforts to measure knowledge flows using scientific literature at different levels of detail, namely: among scientists, among journals, among subject categories, among institutions and among countries.

Zhuge (2006) argues that ideas in a scientific article inspire new ideas, which will be recorded and published as new articles after peer review. Therefore, citations between scientific articles imply a knowledge flow from the authors of the article being cited to the authors of the articles that cite it. Zhou and Leydesdorff (2007) use journal-journal citation analysis to investigate international visibility of journals. Zhou et al. (2010) also use journal-journal citation analysis to study the specialization of a research community within a discipline. Johannes and Guenter (2001) measure knowledge export and international visibility of journals by determining the unique subject fields to which the citing journals have been assigned and the unique countries to which the citing authors belong, respectively.

Rowlands (2002) proposes a method to measure the spread of scientific knowledge that is published in a journal. He focuses on journals as units of spread and introduces an indicator to measure the spread of knowledge by looking at the number of different journals that cite the papers published in the primary journal, as shown in Equation 1.

$$RDI = \frac{U}{Cit}, \tag{1}$$

where U stands for the number journals that cite the papers published in the primary journal in a given time window (say T). Cit is the total number of citations received by the articles in the primary journal in T time window and the notion RDI is for Rowlands Diffusion Index. Naturally, diffusion can only increase in an absolute sense, however, empirical results show that the diffusion index proposed by Rowlands is negatively correlated with the total number of citations received (Rowlands, 2002). This leads Frandsen (2004) to provide a different diffusion index, as shown in Equation 2.

$$FDI = \frac{U}{Pub}, \tag{2}$$

where Pub stands for total number of publications in the primary journal, U is the same as above and FDI stands for Frandsen Diffusion Index. Note that Cit is replaced by Pub (i.e. publications). When publications do not change, the Frandsen Diffusion Index cannot decrease, and thus, the Frandsen Diffusion Index is positively correlated with the total number of citations.

Burrell (1991, 1992, 2005 and 2006) shows that the Leimkuhler Curve can provide an intuitive visual representation for the Gini Coefficient Index in giving graphical and numerical summaries of the concentration of bibliometric distributions. Guan and Ma (2007) illustrate the use of the Leimkuhler Curve to reveal the impacts of research outputs of countries. Using the Gini index, Liu and Rousseau (2010) study knowledge diffusion through publications and citations, as shown in Equation 3.

$$G = \frac{2q - 1}{N}, \text{where}$$

$$q = \sum_{i=1}^{N} i \frac{x_i}{M}$$

$$M = \sum_{i=1}^{N} x_i \tag{3}$$

N denotes the number of subject categories, and x_i denotes number of citations in journals mapped with a given subject category i. Note that the Gini index (Burrell, 1992, 2005) can be equally computed using Equation 4.

$$G = 1 - \frac{\sum_{j=1}^{\infty} (r(j))^2}{N.M}, \quad (4)$$

where M and N are the same as in Equation 3, $r(j)$ stands for the number of subject areas with at least j citations and the sum is finite as there is always a subject category with the largest number of citations. Note that Gini based indexes can only characterize the knowledge diffusion and do not quantify the volume of knowledge flow.

Ingwersen et al. (2000) present international citations as an indicator to measure export of knowledge produced by institutions. They measure knowledge export of institutes by calculating the proportion of citations received by a given institute from other countries (outside the host country where the institute is located) relative to total citations received by the institute. Using citation exchange among the scientific articles, we introduce a notion of International Scholarly Impact of Scientific Research (ISISR) to measure international knowledge flows among countries and institutions (Hassan & Haddawy, 2013). However, the measure of ISISR only quantifies knowledge flows and does not elucidate the contents of knowledge that flows across the countries.

The above survey discusses the salient research to quantitatively measure knowledge flows using bibliometric data. However, we believe that apart from the quantitative measures it is extremely important to analyze the contents of the knowledge flows. The scientific work of Zhuge (2009, 2010, 2011 & 2012) sets the theoretical base of semantic analysis in order to extract knowledge from large scale corpus.

Methodology

This section presents analytical techniques used to semantically analyze the knowledge flow from a given source country. We consider a set of papers $P`$ authored by the researchers of a given source country in a given subject area in a given time window. Among the selected papers, we identify the papers P cited by the papers only authored by researchers from outside the source country. We cluster the papers from P to identify the main topics. We procure the sets of papers (authored by researchers outside the given country) citing each of the topic clusters. Next, we in turn cluster each set of citing papers to again identify main topics in order to identify how the knowledge from the topics in the cited papers is being used. The research topics are identified using our proposed Topic with Distance Matrix (TDM) model, an extension of the Latent Dirichlet Allocation (LDA) model proposed by Blei et al (2003).

A number of approaches to model scientific paper content have been proposed (Blei et al., 2003; Hofmann, 1999). These approaches are based upon the idea that the probability distribution over words in a paper can be expressed as a mixture of topics, where each topic is a probability distribution over words. We utilize one such popular model, LDA, proposed by Blei et al. (2003). In LDA, the generation of a paper collection is modeled as a three step process. First, for a given paper, a distribution over topics is sampled from a Dirichlet distribution. Then, for each word in the paper, a single topic is selected according to this distribution. At Last, each word is sampled from a multinomial distribution over words specific to the sampled topic.

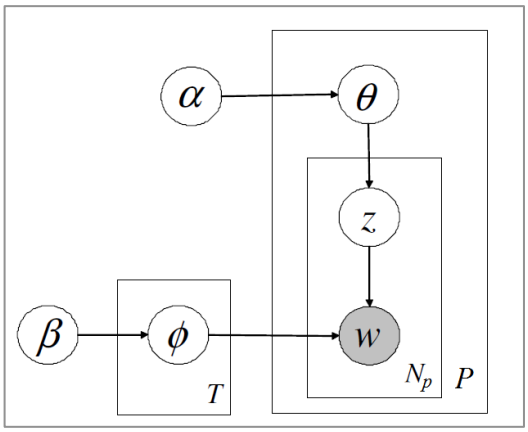

Figure 1. Latent Dirichlet Allocation (LDA) Model.

Using plate notation, the generative process corresponding to the hierarchical Bayesian model is shown in Figure 1. In this model, Φ stands for the matrix of topic distributions for each of T topics being selected independently from a symmetric Dirichlet prior (β). Θ is the matrix of paper specific mixture weights for these T topics, each being drawn independently from a symmetric Dirichlet prior (α). For each word, z denotes the topic responsible for generating that word, drawn from the Θ distribution for that paper, and w is the word itself, drawn from the topic distribution Φ corresponding to z. A paper p is a vector of N_p words, w_d, where each w_{id} is chosen from a vocabulary of size V and P is a collection of papers.

Estimating Θ and Φ provides information about the topics that participate in a publication corpus and the weights of those topics in each paper respectively. A variety of algorithms have been used to estimate these parameters, including variational inference (Blei et al., 2003), expectation propagation (Minka & Lafferty, 2002), and Gibbs sampling (Griffiths & Steyvers, 2004). To induce the probability distribution of Θ and Φ, LDA uses Gibbs Sampling which starts from randomly selected initial states and then revises distributions by changing topics to find correct distributions. Finally, the model provides topic-word relationship by the vector formed probabilistic representations.

Using the LDA, we obtain topic vectors where each value in the vector is associated with a given word that shows the probability of the word occurring under the given topic. For instance, vector T_1 ($word_1$: 0.3, $word_2$: 0.1, $word_3$: 0.2, …, $word_n$: 0.8) shows the probability distribution of all n words for the given topic t_1. Using this information, we represent each paper (from the set P) in the form of a vector where each value in the vector represents the probability distribution of a given word from vocabulary V in the paper for the topic under consideration (say t_1). For instance, P_1 ($word_1$: 0.4, $word_2$: 0.2, $word_3$: 0.0, …, $word_n$: 0.7) shows the probability distribution of words in the paper p_1 for the topic t_1. Note that if a word from V does not appear in p_1 then we assign default zero probability for that word.

Using the Minkowski distance between a given paper-vector P and topic-vector T, we choose papers in order to classify them as belonging to a specific topic (see Equation 5).

$$D = \sqrt{\sum_{i=1}^{n} | a_i - t_i |^2}, \tag{5}$$

where a_i denotes the probability of the term i in paper p_1 for the given topic T, and t_i denotes the probability of term i for the topic T. In order to obtain a set of papers relevant to topic T, a threshold TH is applied with the given percentage of the distance between the minimum and the maximum distance of paper vectors from T. Our experimental results show that the highest F-measure is achieved with TH = 25%. The size of a topic is determined by the

number of papers associated with it. The numbers of topics are determined by computing inter and intra topic similarity. We minimize inter topic similarity and maximize intra topic similarity to obtain the optimal number of topics. To compute the inter similarity between two topic, we use the Jaccard distance index (Jaccard, 1901).

Case Study: Semantic Analysis of Knowledge Flows across Countries in the Field of Renewable Energy

Dataset

We present a case study to illustrate the use of our technique in the subject area Renewable Energy. Using All Science Journal Classification (ASJC), we procured 46,518 publications (journal articles, reviews and conference papers) classified as Renewable Energy, a subarea of Energy(all) from the Scopus database during the time period 2005-2010

We procure 8,590 papers ($P`$) (journal articles, reviews and conference papers) published by researchers from the United States. Among the selected set of papers $P`$, we select 4,362 papers (P) which are cited by papers authored only by researchers from other countries. Further, we select candidate terms to represent each paper. In order to procure such terms, we use author defined keywords from the selected papers. In addition, we extract noun terms from the abstracts and titles of the papers using SharpNLP (http://www.codeplex.com/sharpnlp). We then identify synonyms of the selected noun terms using WordNet 3.0 (http://wordnet.princeton.edu/) and include them as candidate terms as well. Next, we apply the Porter Stemming algorithm (http://tartarus.org/martin/PorterStemmer/) to stem all the selected candidate terms. Finally, we feed this data to our TDM model.

Research Topics Cited by Researchers from Outside the United States in the Field of Renewable Energy

Figure 2 shows four research topics in the field of Renewable Energy cited by researchers from outside the United States. Using Wordle.Net (http://www.wordle.net/), we visualize the contents in each topic. Here, each topic is represented with the most frequently occurring author defined keywords collected from the papers in a given topic. The number of papers belonging to a specific research topic and the size of each research topic are written next to its respective topic. The research topics 1 and 4 are the largest topics cited by researchers from outside the United States. The topic#1 is the largest topic, containing 44% of the 4,362 papers. This topic covers research work related to Solar Cells, Solid Oxide Fuel Cells (SOFC) and Proton Exchange Membrane Fuel Cells (PEMFC). The topic#2 is related to Hydrogen Production. This topic also covers research related to Steam Reforming, a method for producing hydrogen, carbon monoxide, or other useful products from hydrocarbon fuels such as natural gas. Finally, the topic#3 is about Li-ion batteries. Li-ion batteries are an important type of rechargeable battery, particularly used in mobile devices. Finally, the topic#4 covers research related to Sustainable Management. Next we explore how the researcher from different countries cites the knowledge produced by the United States.

Research Topics of the Publications Produced by Chinese and Japanese Researchers that Cite Papers Authored by Researchers from the United States

To understand the difference in the use of the same knowledge, we further analyse that how the scientific knowledge diffuses into other research topics used by different research communities. We compare publications of the researchers from China and Japan that cite the same knowledge produced by the researchers from the United States. We select topic#1 from Figure 2 (the largest topic cited by the researchers from outside the United States in the field of Renewable Energy during 2005-2010). This topic covers research topics related to Solar

Cells (including Thin Film Solar Cells, Solid Oxide Fuel Cells and Proton Exchange Membrane Fuel Cells). Furthermore, we procure all the papers (journal articles, reviews and conference papers) authored by researchers from China and Japan that cite papers in the selected topic. We then identify research topics of the selected Chinese authored and Japanese authored papers.

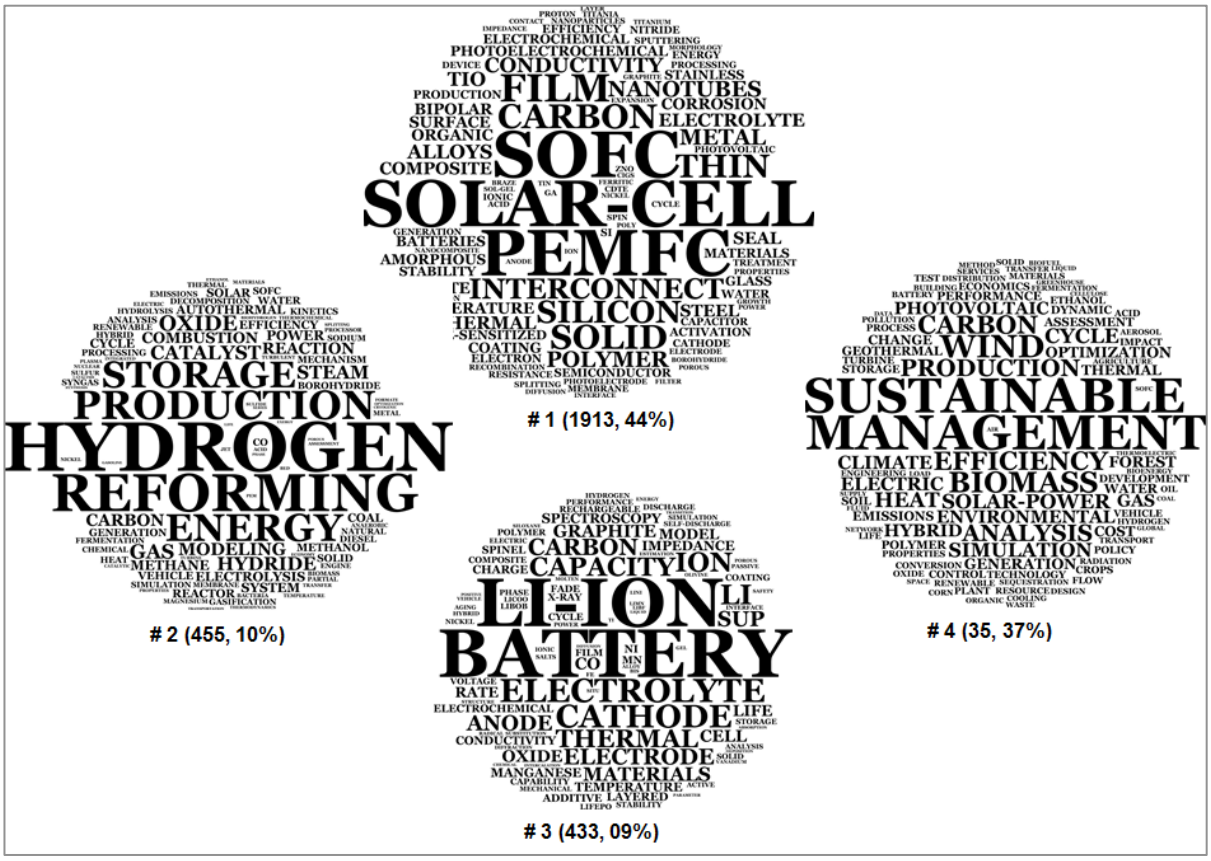

Figure 2. Research Topics Cited by Outside the United States in the Field of Renewable Energy during 2005-2010.

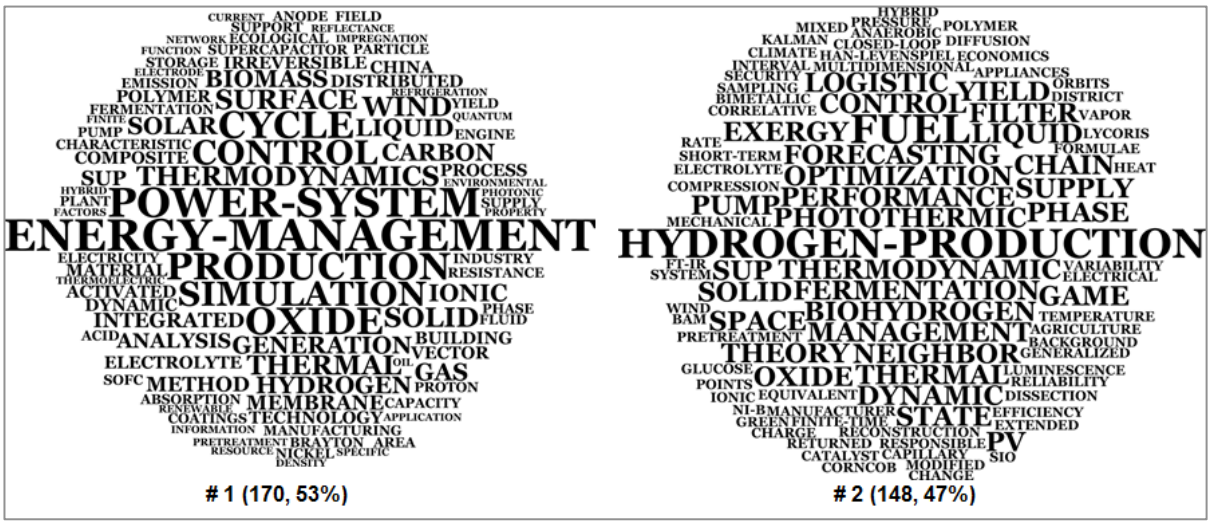

Figure 3. Research Topics of the Scientific Knowledge Produced by the Chinese Researchers (during 2005-2010) that cite the topic#1 in Figure 2.

Figure 3 shows research topics of the scientific knowledge produced by the Chinese researchers during 2005-2010 that cite topic#1 in Figure 2. In Figure 3, topic#1 mainly covers research related to Power Systems, Energy Management and Production. This topic is the largest topic which contains 53% papers out of 318. The topic#2 which contains 47% of the papers mainly focuses on Hydrogen Production.

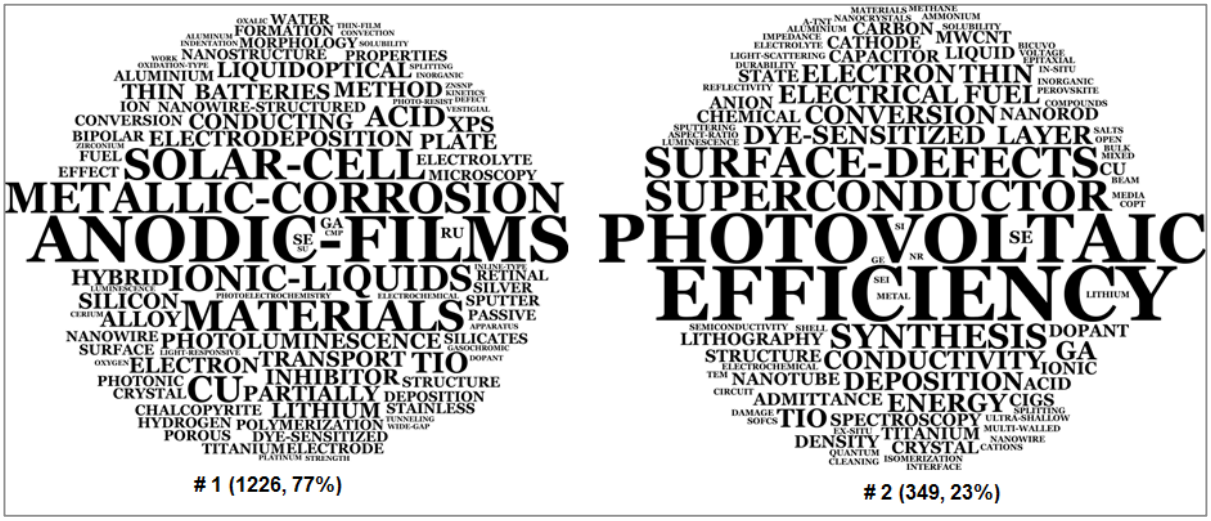

Figure 4. Research Topics of the Scientific Knowledge Produced by the Japanese Researchers (during 2005-2010) that cite topic#1 in Figure 2.

Figure 4 shows research topics of the scientific knowledge produced by the Japanese researchers during 2005-2010 that cite topic#1 in Figure 2. In contrast with China, the Japanese research community utilizes the same knowledge (produced by the United States) in rather different research themes. The Japanese researchers focus on topics related to Metallic Corrosion and Anodic Oxide Films (see topic#1 in Figure 4). Interestingly, we also find another topic (topic#2: 55 papers) describing the efficient use of Photovoltaics, Dye-sensitized Solar Cells and Superconductors. Note that Superconductors play a vital role in providing low-cost renewable energy.

Concluding Remarks

In this paper we have presented a new topic model with distance matrix, called TDM, to semantically analyze knowledge flows across countries by using publication and citation data. We have also presented a case study to illustrate the use of our proposed techniques in the subject area of Renewable Energy during 2005-2010 using data from the Scopus database. We have compared the Japanese and Chinese papers that cite the same scientific literature produced by the researchers from the United States in order to show the difference in the use of same knowledge. The study has shown that Japanese researchers focus in research areas such as efficient use of Photovoltaics, and Superconductors (to produce low-cost renewable energy). In contrast with the Japanese researchers, Chinese researchers focus in the areas of Power Systems, Power Management and Hydrogen Production.

The method of semantic analysis presented in this paper provides an understanding of the internationality of research not provided by studies of researcher mobility and co-authorship patterns. Our case study highlights the diversity in the ways that research produced by a country may be used in different international contexts, even within a relatively narrow research area. Such analyses may be helpful in establishing more effective multi-national research collaboration and in aligning collaboration with national priorities.

References

Blei, M., Ng, A. & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Burrell, Q. L. (1991). The Bradford distribution and the Gini index. *Scientometrics, 21*, 181–194.

Burrell, Q.L. (1992). The Gini index and the Leimkuhler curve for bibliometric processes. *Information Processing and Management, 28*(1), 19-33.

Burrell, Q.L. (2005). Measuring similarity of concentration between different informetric distributions: Two new approaches. *Journal of the American Society for Information Science and Technology, 56*(7), 704-714.

Burrell, Q.L. (2006). Measuring concentration within and co-concentration between informetric distributions: An empirical study. *Scientometrics, 68*(3), 441-456.

Frandsen, T. (2004) Journal diffusion factors: A measure of diffusion?. *Aslib Proceedings, 56*(1), 5-11.

Griffiths, T. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(1), 5228–5235.

Guan, J. & Ma, N. (2007). A bibliometric study of China's semiconductor literature compared with some other major Asian countries, *Scientometrics, 70*(1), 107-124.

Hagel, J., Brown, J. & Davison, L. (2009). *Measuring the forces of long-term change: The 2009 shift index*. Deloitte Development LLC.

Hassan, S. & Haddawy, P. (2013). Measuring international knowledge flows and scholarly impact of scientific research, *Scientometrics*, *94*(1), 163–179.

Ingwersen, P., Larsen, B. &. Wormell, I. (2000). Applying diachronic citation analysis to ongoing research program evaluations. *In B. Cronin & H.B. Atkins (Ed.), The Web of Knowledge (pp. 373-387)*. Medford, N.J.: Information Today, Inc. & American Society for Information Science.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura, *Bulletin de la Société Vaudoise des Sciences Naturelles, 37*, 547–579.

Johannes, S. & Guenter, G. (2001). Citation rates, knowledge export and international visibility of dermatology journals listed and not listed in the Journal Citation Reports. *Scientometrics, 50*(3), 483-502.

Liu, Y. & Rousseau, R. (2010). Knowledge diffusion through publications and citations: A case study using eSI-fields as unit of diffusion. *Journal of the American Society for Information Science and Technology, 61*(2), 340-351.

Minka, T., & Lafferty, J. (2002). Expectation-propagation for the generative aspect model. *Proceedings of the Eighteenth Conf. on Uncertainty in Artificial Intelligence*, 352–359.

Rowlands, I. (2002). Journal diffusion factor: A new approach to measuring research influence. *Aslib Proceedings, 54*(2), 77–84.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic Author-Topic Models for Information Discovery. *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, Washington.

Zhou, P. & Leydesdorff, L. (2007). A comparison between the China scientific and technical papers and citations database and the Science Citation Index in terms of journal hierarchies and inter-journal citation relations. *Journal of the American Society for Information Science and Technology, 58*(2), 223-236.

Zhou, P., Su, X., & Leydesdorff, L. (2010). A comparative study on communication structures of Chinese journals in the social sciences. *Journal of the American Society for Information Science and Technology, 61*(7), 1360-1376.

Zhou, P. & Leydesdorff, L. (2007). A comparison between the China scientific and technical papers and citations database and the Science Citation Index in terms of journal hierarchies and inter-journal citation relations. *Journal of the American Society for Information Science and Technology, 58*(2), 223-236.

Zhou, P., Su, X. & Leydesdorff, L. (2010). A comparative study on communication structures of Chinese journals in the social sciences. *Journal of the American Society for Information Science and Technology, 61*(7), 1360-1376.

Zhuge, H. (2006). Discovery of knowledge flow in science. *Communications of the ACM, 89*(5), 101-107.

Zhuge, H. (2009). Communities and Emerging Semantics in Semantic Link Network: Discovery and Learning, *IEEE Transactions on Knowledge and Data Engineering*, *21*(6), 785-799.

Zhuge, H. (2010). Interactive Semantics, *Artificial Intelligence*, *174*, 190-204.

Zhuge, H. (2011). Semantic linking through spaces for cyber-physical-socio intelligence: A methodology, *Artificial Intelligence*, *175*, 988-1019.

Zhuge, H. (2012). Knowledge Flow, Chapter 5 in The Knowledge Grid - Toward Cyber-Physical Society, 2nd Edition, *World Scientific Publishing Co.*, Singapore.