Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 1, pp. 42-51

Mining Authoritativeness of Collaborative Innovation Partners

J. Engler, A. Kusiak

Joseph Engler
Adaptive Systems Rockwell Collins, Inc.
Cedar Rapids, IA 52498 USA
E-mail cjengler1@msn.com

Andrew Kusiak
Mechanical and Industrial Engineering
3131 Seamans Center University of Iowa
Iowa City, IA 52242-1527 USA
E-mail: andrew-kusiak@uiowa.edu

Abstract:
The global marketplace over the past decade has called for innovative products and cost reduction. This perplexing duality has led companies to seek external collaborations to effectively deliver innovative products to market. External collaboration often leads to innovation at reduced research and development expenditure. This is especially true of companies which find the most authoritative entity (usually a company or even a person) to work with. Authoritativeness accelerates development and research-to-product transformation due to the inherent knowledge of the authoritative entity. This paper offers a novel approach to automatically determine the authoritativeness of entities for collaboration. This approach automatically discovers an authoritative entity in a domain of interest. The methodology presented utilizes web mining, text mining, and generation of an authoritativeness metric. The concepts discussed in the paper are illustrated with a case study of mining the authoritativeness of collaboration partners for microelectromechanical systems (MEMS).
Keywords: Innovation, web mining, text mining.

1 Introduction

Innovation most often occurs in one of two forms, incremental or radical. Radical (discontinuous) innovation assumes focuses on a completely new concept that is radically different from the existing ideas. This type of innovation occurs rarely and is not easily predictable. Incremental, or continuous, innovation builds upon previous concepts and therefore it is easier to be quantified. Kusiak [1] defined innovation as an iterative process aimed at the creation of new products, processes, knowledge or services by the use of new or even existing technology. This definition summarizes the typology of incremental innovation.

Innovation has further been quantified into five generational models. The first generational model is linear, where innovation is unidirectionally pushed from the research phase to the commercial application phase [2], [3]. The second model, the pull model, holds the consumer as the main focus of innovation as opposed to the designer [4]. Feedback forms the third model and utilizes the consumer's responses to an initial product/service offering to perform incremental innovation on that product and/or service [5]. The fourth model is known as the strategic model in which innovation lines up directly with the company's strategy [2].

The model pursued in this paper is the fifth model, also known as the networked model. In this model extra-enterprise and cross-discipline organizations form a network to innovate. The term Open Innovation [6] is often used when describing this model. Collaborative networking involves a detection of the optimal sources of collaboration. This can often be viewed as a challenge and often results in the local optimum as the choice of the collaborative source rather than the unseen, and most often unknown, global optima (e.g., the most authoritative person/company).

The success of collaborative innovation depends greatly on the quality of the collaboration sources. The probability of innovation success, as measured by the market, can be considered proportional to the quality of the collaborative sources. Therefore, it is incumbent upon a company, in pursuit of innovation, to seek out the optimal sources for collaboration. A wealth of information is available upon the World Wide Web (WWW) for identifying the optimal sources of collaboration (e.g., white-papers written by authorities of specific domains).

Various researchers have begun to investigate possible means by which collaboration produces effective results. Unfortunately, the literature is lacking solid systematic methodologies by which collaboration authoritativeness may be determined apriori of collaboration inception. Chapman et al. [22] proposed a process model for collaborative data mining in an electronic manner but fail to address the need for authoritativeness when determining the collaborative partner selection process. Lavraĉ et al. [23] surveyed various methodologies for collaboration but fail to report a systematic methodology of determining the most authoritative entity for collaboration. Gajda [24] investigated assessment measures for determining the success of a collaborative partnership on a project. Unfortunately, these measures are considered upon the completion of the project rather than to determine the effective partners for collaboration prior to entering in to a collaborative agreement.

Other researchers have posited criteria and methodologies for collaboration partner selection. Geringer [25] proposed task-related selection criteria for international joint ventures but failed to present a systematic methodology for automatically determining the authoritativeness of a collaborative partner. Hitt et al. [26] investigated resource-based and organizational learning for collaborative partner selection. Again, the methodologies in the literature fall short of the goal of automated systematic determination of authoritativeness for collaboration.

This paper presents machine learning algorithms to extract collaborative innovation relationship information from various sources including the WWW. Utilizing machine learning algorithms to discover valuable knowledge from disparate sources has been presented in the literature. Chen [27] posited utilizing natural computing techniques, such as swarm intelligence, to foster a collective intelligence in a virtual learning environment. Grebla et al. [28] presented a Bayesian belief network for mining data from various databases to assist in predicting arteriosclerosis and cardiovascular disease.

This paper offers a novel methodology by which the optimal authoritative sources of collaborative partnership may be discovered. Through the use of web mining, text mining, and the creation of an authoritativeness matrix, users may determine the optimal authority with which to perform collaborative innovation. Of course, optimality may depend upon more than just the most authoritative partner on a given subject. Other factors such as the availability of the collaborator or the cultural background of the collaborator (e.g., defense systems collaboration) may be involved. Thus, this paper advocates the creation of an authoritativeness matrix as opposed to simply defining global optima for collaboration.

The remainder of this paper proceeds as follows. Section 2 discusses the focused mining of the World Wide Web to discover authoritative sources for collaboration. The distillation of these sources to form an authoritativeness matrix is discussed in Section 3. An authoritativeness metric is presented in Section 4. Section 5 discusses clustering of the mined sources of collaboration. Section 6 offers a case study for determining the leading authorities on microelectromechanical systems (MEMS). Finally, Section 7 offers concluding remarks.

2 Focused Mining of the Web

The first major step in forming a collaborative innovation relationship is to seek out and choose partners for the collaboration process. The World Wide Web (WWW) presents a proven search space for multiple concepts. A natural inclination is to manually search the internet for such sources of collaboration. Some companies hire business development teams to perform this task. Manually searching the web is a time intense process that often yields sub-optimal results. Sometimes searches can even present misguided or influenced results due to the ability for parties to influence their rank among the various search engines [7].

Increasing the difficulty of the manual search, many search engines utilize payer based rankings which facilitate assigning a higher position in the search results. Additionally, many websites have multiple internal links thus boosting certain search engine ratings. Such forms of ranking manipulation may provide false results and thus the most suitable collaborative candidates could be missed.

To overcome the limitations of a manual search of the internet for collaboration resources, a focused web miner is presented. The focused web miner used for this process includes user inputs of a specified phrase which become the search criteria. The focused web miner then proceeds, following standard focused web crawling methodologies as presented by Liu [8], to traverse the WWW in search of white papers, articles, and journal entries related to the search criteria. The presented web miner can easily be extended to handle other sources, e.g., information about companies which have reached Phase II funding from Small Business Innovation Research (SBIR) programs.

The presented version of the focused web miner does not attempt to mine the web blog data or the standard html files. Rather, this version of the focused web miner seeks content mimicking academic writings. Thus, the focused web miner spends a fair amount of time searching academic web sites, scientific communities, and trade journals. It is from these types of internet resources that the, often academic, writings are extracted.

The discovery of web pages containing white papers is a significant task involving crawling the internet and classifying the web pages that are examined as a review or non-review page. The standard approach to performing the crawl is to utilize a focused web crawler. A focused web crawler targets a specific corpus of web pages. Standard crawling does not consider a specific topic of inquiry; rather its job is to index all pages available on the internet.

Even with the assistance of algorithms such as PageRank developed by Google [17], successful standard crawling requires massive hardware and bandwidth. This drawback prevents most corporations from performing this type of crawl internally. Focused crawling requires far less hardware and bandwidth but does require some sophistication of algorithms to weed out the undesirable links as they relate to the given query. The algorithms useful to focused web crawling involve basic classification algorithms.

There is a great variety of classification algorithms for determining web pages containing white papers. Shih et al. [18] suggested the use of web page content structure as parameters of classification. The authors in [18] indicated that content providers tend to choose URLs and page layouts that coherently structure their content. This html structure may be useful in determining the likelihood that a web page contains reviews. Kules et al. [19] extended this idea to limit the features used for classification to items such as web page titles, URLs and text snippets. Jin et al. [20] took a different approach from the previous two and utilized a data-mining algorithm called Hidden Nad'ye Bayes. Their methodology considered a large corpus of web pages and calculated the probability that each page fit into a particular category.

Fortunately for the focused web miner proposed here, document type is often the best indicator of a possible fit. Given the query string provided by the user and a set of allowable document types (e.g., pdf files only) the need for classification is reduced greatly. The need for classification increases dramatically when white papers other than standard academic content is searched. In such case the use of a combination of Bayesian classification and the web page structure algorithm of [18] is suggested (see [21]).

Once the focused web miner has discovered a document of the standard academic writing, it attempts to download this document to a central repository that is dedicated to the search criteria. This repository often becomes extremely large, in the range of a terabyte or more, but is central to the process for collaboration resource detection. No partitioning of this repository takes place at the time of download; rather, all documents are placed into the same location. It is upon these documents that the process of determining the most authoritative collaboration partners is performed. While this form of repository may seem excessive, the reader is reminded of the low cost of data storage. Additionally, a great deal of information will be gleaned from this repository over time.

3 Authoritativeness Matrix

An authoritativeness matrix is generated from the documents that were obtained from the focused web mining process. Utilizing standard text mining techniques the documents are deconstructed to gain the information necessary for the generation of the authoritativeness matrix. A text-mining algorithm extracts from each document the author's names and the references, other authors, cited by that particular paper.

To assist in discovery of the authors who wrote, and are cited in, the papers a list of first names is utilized. The list of first names, freely available on many internet sites, allows for the detection of document patterns within the corpus such that most often names of authors are placed within a given context within the document (e.g., author names at the beginning, authors who are cited at the end). The text mining algorithm utilizes these patterns to classify portions of the document which will have the information extracted from. Additional, text-mining algorithms may be utilized to detect the sense, positive or negative, in which a citation appears within the document.

From the information mined by the text-mining algorithm the authoritativeness matrix may be constructed. The authoritativeness matrix is a two dimensional matrix, or table, made up of columns representing individuals who have been referenced by the papers and rows representing paper authors. The authoritativeness matrix forms a concise but sparsely populated representation of the given, or presented, authorities in the documents.

Figure 1 presents an example authoritativeness matrix for three documents. The rows represent the authors of the documents while the columns represent authorities who are cited as references in these documents. A cell of the matrix is 1, if the author in the row containing the cell has referenced an authority in the column containing the cell. Otherwise the cell is 0. Each row represents a single author for a single paper, thus there may exist multiple rows for a single paper.

From the example of Figure 1 it is easily determined that JR Koza is the most authoritative person with which to conduct collaborative innovation for the example domain. This is due to the fact that JR Koza is the most cited

	Howit, P.	Benkler, Y.	Stokic, D.	Von Hippel, E.	Nolan, R.L.	Hansen M.T.	Koza, J.R.	Document year
Kusiak, A.	1	1	0	1	1	0	1	2006
Lin, G.	0	1	0	0	1	1	1	2004
Stokic, D.	1	0	1	0	0	0	1	1999

Figure 1: Example authoritativeness matrix for three documents

author within the tiny corpus of documents for the example. It will be demonstrated later in this section that other factors contribute to this outcome.

It should be noted that an author may reference his/her own writings as well as be referenced by others. Thus, it is possible for this matrix to hold entities that are in both the rows and columns of the matrix. In fact, it is possible, although highly improbable that the authoritativeness matrix holds exactly the same authors in its rows as it does references in its columns. As will be explained in Section 4, there some caution needs to be exercised when an author often references their own work while others do not. This caution is the motivation for the storage of the document year in the authoritativeness matrix as will be explained.

The representation of the authorities of the documents discovered by the focused web miner by the authoritativeness matrix ensures ease of storage and traversal. The authoritativeness matrix is compact enough to be stored in main memory especially given the sparseness of the matrix. This allows for efficient processing when determining the true authorities for the collaboration process as presented next.

4 Authoritativeness Metric

To determine which entities in the authoritativeness matrix represent the optimal, or most, authoritative entity, within the search criteria, an authoritativeness metric is used. This metric accounts for the number of documents which were written by the authority, the number of times the authority is referenced in the body of work, discovered by the focused web miner, as well as the average age of the documents written by the authority. Additionally, the sense, positive or negative, in which the author is portrayed in the document can be collected.

Thus, the first step in calculating the authoritativeness metric is to scan the authoritativeness matrix and calculate a number of measures. These measures will be stored in various hash tables for efficient referencing. During the scan of the authoritativeness matrix a hash table representing the authors, or rows, of the matrix is created. Each time an author is encountered in the scan of the matrix that author is either added to the hash table as a key value pair of <AuthorName, 1> or the value of the index of the author in the hash table is incremented. The same action is performed for the columns, or referenced authorities, in the matrix. Additionally, a hash table is created for the purpose of obtaining the average age of the documents for each author, or row, in the matrix. The use of the three hash tables makes for an efficient scanning methodology for the authoritativeness matrix since the matrix is scanned only a single time. A forth table may be required to represent sense.

The hash table which represents the number of documents written by the authors, or rows, of the authoritativeness matrix is used to obtain the number of out-links of each author. Out-links are documents written by the author. Similarly, the hash representing the number of times an author is referenced is used to obtain the number of in-links for each authority. In-links are documents written by other authors referencing the given author, thus implicitly conveying authority on, or detracting from in the case of a negative sense, the referenced author.

The conveyance of authority on an author by referencing their work is held here in a context similar to that of the PageRank algorithm discussed in [9]. Conveyance of authority plays a vital role in the determination of authoritative collaborative sources. Thus, the authoritativeness metric proposed weights the in-link measure higher than the out-link measure. Additionally, should sense be included, the in-link could have the ability to decrease the author's authoritativeness.

The initial authoritativeness metric is defined in (1). In (1), λ is a user defined parameter which allows for

changing the weight of the out-links based on the average age of the documents written by author a_i. This parameter assists in controlling the conveyance of authority to an author who is a prolific writer but perhaps not often cited by other authors. Additionally, the λ parameter allows for decreasing the weight of older documents if only recent documents are desired. If sense is utilized, the in-link, in_i, can be a negative number. It should be noted that determination of the final authoritativeness metric measure is an iterative process as will be explained.

$$A_i = \ln(\lambda^{t'}(out_i) + in_i) \tag{1}$$

where:
out_i is the number of out-links;
in_i is the number of in-links;
t' is the average age of the documents of author a_i in years from the current year;
λ is a user parameter in the range of [0, 1].

Thus, the initial authoritativeness of an author ai is given by A_i. Once the initial individual authoritativeness metrics of the entire authoritativeness matrix is calculated, the iterative process of boosting the authoritativeness is performed. Similar to the methodology used in the PageRank algorithm [10], it is desirable to instill more authority to an author who is referenced by another author of high authority. Thus, if one author is considered a leading authority, deemed so by the authoritativeness metric, and that author references a second author, the second author's authoritativeness metric measure should be increased.

The iterative process of authoritativeness boosting is performed using the average of the in-links pointing to the current author. In-links of high authority contribute to the boosting of the current author's authoritativeness, while in-links of less authoritative authors are not detrimental to the current author's authoritativeness. Thus the average of the in-links during the authoritativeness boosting process is calculated by including the authoritativeness of the author who referenced the current author as shown in (2).

$$\overline{in_i} = \sum_{i=1}^{N} \frac{e^A}{N} \tag{2}$$

where:
A_i is the authoritativeness of author a_i;
N is the number of in-links to a_i.

At each iteration of the calculation of the final authoritativeness metric the average of the in-links is used to calculate the new authoritativeness metric measure A_i for each author a_i. Equation (3) describes how the new in-link measure is calculated iteratively.

$$in_i^{'} = \sum_{j-1}^{N} \begin{cases} e^A - \overline{in_i} & \text{if } e^A > \overline{in_i} \\ 1 & \text{if } e^A \leq \overline{in_i} \end{cases} \tag{3}$$

Thus, at each iteration of the boosting process, ini of (1) is replaced with in_i for the calculation of the authoritativeness of each author. The boosting process is continued for n iterations, as set by the user, or until the order of the authorities remains unchanged which is the preferred method.

With the final authoritativeness metric at hand for each author it is easy to determine which author and/or entity is the most authoritative in the subject matter of the search criteria. It is useful to determine the top k authorities in the subject matter to ensure that a good collaborative resource can be found available and willing to collaborate on the innovation at hand. As such it is useful to set a user defined parameter λ which is a threshold below which authoritativeness is discounted. This threshold is utilized in determining the authorities of the clustered documents as explained next.

5 Document Clustering

Often the size of the search space or the generality of the search criteria can result in a document set of varied type which is large in size. To ensure that the collaborative partner chosen by the authoritativeness metric is the one

that is most appropriate for the specific collaboration it is helpful to cluster the documents into similar categories. Once the clustering has been performed, a cluster that is most similar to, or most represents, the specific innovation topic is chosen. From that cluster it is the possible to determine the best collaborative source for the innovation. Note, the collaborative authority of a specific cluster may not be the authority whose overall authoritativeness metric is the highest. Rather, the cluster authority, or authorities, will be those who are most advantageous for the specific innovation subject.

Clustering of the documents mined via the focused crawler begins with the generation of a word frequency matrix for the documents. The word frequency matrix represents the counts of each word in the individual documents. Each row of the matrix represents a single document; while each column of the matrix represents a single word. There exists columns for every document word, which is not a stop word, thus the matrix can be somewhat sparse. Many words, known as stop words, do not assist in properly classifying the documents. Stop words are most common in everyday language and thus not specific to the topic. Words such as "the", "in" and "here" are removed from the word frequency matrix prior to clustering. Further, it is often favorable to generate the root of words as opposed to the actual words for this frequency matrix. Thus, words such as "innovation", "innovate", and "innovativeness" would all be placed in the root word frequency cell for the word "innov". Figure 2 below represents a partial frequency word matrix.

	Ability	Able	Absolute	Abstract	Accelerate	Accept
00006009.pdf	1	5	0	0	0	4
00104286.pdf	0	0	0	0	0	0
00183750.pdf	2	0	2	1	0	0
00263768.pdf	1	0	2	0	0	0
00267882.pdf	0	0	0	1	0	0
00522535.pdf	1	0	0	1	21	0
00540568.pdf	0	0	1	1	0	0
00608091.pdf	0	0	0	0	1	1
00653292.pdf	0	1	0	1	0	2

Figure 2: Word frequency matrix

Once the word frequency matrix is obtained it is important to reduce the dimension of the matrix to ensure efficient clustering. Dimensionality reduction techniques, such as singular value decomposition, that are used for standard data mining are especially helpful here. The word frequency matrix before dimensionality reduction can easily include thousands of words or attributes. Rarely are all the attributes of value to the clustering. Thus, by performing a dimensionality technique such as singular value decomposition, the attribute set can be reduced down to a size that is more manageable, typically of size 100 or less [11].

Once the dimensionality reduction has been performed, the reduced word frequency matrix is clustered with simple k-means clustering algorithm described in [11]. Thus, a brief review of the cluster centroids will help to determine which cluster most resembles the subject matter of the specified innovation.

The authoritative collaboration partner(s) can easily be determined from those entities that have contributed work to the cluster that most resembles the subject of the innovation. Section 4 presented a threshold measure by which authorities could be weeded out of the collaborative search process. Following the Apriori Property discussed in [11] and [12], those authorities that are not authoritative for the entire group should not be considered authoritative for a subsection of that group. Therefore, only authorities with the authoritativeness metric higher than the user defined threshold should be sought within the clusters.

6 MEMS Case Study

This section presents a case study on the discovery of the most authoritative person to perform collaborative innovation with for the domain of microelectromechanical systems (MEMS). In this study 2403 papers were mined from the internet on the subject of MEMS

Simon [13] describes MEMS as a monolithically integrated device used for microwave applications such as switches, distributed phase shifters and BPSK modulators. Other applications for MEMS have also surfaced. In fact, according to Maeda et al. [14] MEMS is expected to be one of the most promising areas of research and development contributing to future success of electronics businesses.

After the 2403 papers were mined from the internet using the focused web miner, the author's names and references were parsed from the documents as described in Section 3 above and the authoritativeness matrix was generated. The authoritativeness metric, described in Section 4, was applied with λ set to 0.80 to slightly discount the average age of the documents. Figure 3 illustrates the top ten authorities after this initial calculation of authoritativeness. Figure 3 illustrates the results of running the algorithms presented in this paper prior to the iterative boosting discussed in Section 4. Thus, the results in Figure 3 are more indicative of a rapid manual search.

Author	Authoritativeness Metric
GM Rebeiz	4.5218
S Eshelman	4.1897
CL Goldsmith	4.0073
R Langer	3.8286
TA Desai	3.8067
A Malczewski	3.7136
MJ Cima	3.5835
DJ Beebe	3.5553
B Pillans	3.5553
J Ehmke	3.5264
JB Muldavin	3.4965

Figure 3: The top 10 authors in the non-boosted authoritativeness metric list for MEMS

As seen in Figure 3, GM Rebeiz is indicated as the leading authority on MEMS. A quick search of the internet with the name GM Rebeiz justifies his rank as the top authority in this non-boosted list. GM Rebeiz is a professor at the University of Michigan in the College of Engineering and leads a team of 8 PhD students in a focus on RF-MEMS [15].

Utilizing the authoritativeness matrix, discovered in the mining of the 2403 documents which generated the non-boosted results of Figure 3, the iterative boosting of authoritativeness is applied. Upon the application of the boosting of the authoritativeness the list changes in order as can be seen in Figure 4. Boosting has the effect of attributing higher authority to those whose papers have been cited by authors of higher authority. Thus, this is the list that a person for collaborative practices in the field of MEMS should be sought from.

Author	Authoritativeness Metric
CL Goldsmith	19.1892
M Sarantos	18.1891
H Fudem	18.1891
B Pillans	18.1849
S Eshelman	17.9906
RF Lohr	17.6873
D Strack	17.67790
F Kuss	17.6779
E Niehenke	17.6779
E A Sovero	17.4917

Figure 4: Top 10 authors in the boosted authoritativeness metric list for MEMS

From the boosted authoritativeness it is easy to see that CL Goldsmith is the authoritative figure one would wish to collaborate with. In fact, with a quick search of the internet it is found that CL Goldsmith is the president of a company called Memtronics and received his PhD from the University of Texas [16]. The list contains other potential candidates who may be sought after should CL Goldsmith not be available for collaboration.

For this case study, the focused web miner ran for approximately 18 hours to gather the 2403 documents. The parsing of the author's references and document age took less than 2 minutes. The initial authoritativeness was then calculated from the matrix in approximately 1.5 minutes. The boosting of the authoritativeness took 16 iterations

before order no longer changed and took less than 10 minutes to achieve (see Figure 5). Thus, overall, the process of mining the leading authority in the field of MEMS, based upon these documents, took less than 18.5 hours and required very little of the user's time to perform. It is easily seen that this is a marked improvement upon a manual search.

Activity	Time/Quantity
White papers mined	2403
Focused miner runtime	18 hours
Parse authors, references, age	2 minutes
Initial authoritativeness calculation	90 seconds
Number of boosting iterations	16
Boosted authoritativeness calculation	10 minutes
Total algorithm run time	18 hours 13.5 minutes

Figure 5: Summary of the algorithm run

The case study illustrates the effectiveness of the authoritativeness metric presented in this paper. Further, the case study highlights the differences between boosted and non-boosted authoritativeness. From the perspective of a company that is seeking a collaborative partner, the boosted authoritativeness offers a list of highly respected candidates. Deriving this list in the short unmanned time frame of 18.5 hours offers companies a great benefit in discovering the most authoritative person(s) to perform collaborative innovation with.

The effectiveness of the authoritativeness metric is further explained in a recent scenario that was encountered by an electronics manufacturer who required expertise with legacy 16 bit PCMCIA PC cards. Due to confidentiality the details of this episode cannot be related although a summary of the scenario can be provided. The electronics manufacturer, a government contractor, was contracted to design a laptop integrated testing device for a piece of electronic equipment for a foreign concern. One of the requirements for this testing device was for it to integrate with the laptop through a legacy 16 bit PCMCIA PC card. The contractor lacked the domain knowledge to effectively and rapidly design the testing device with this form of legacy interface. Therefore, the author's were asked to apply the authoritativeness metric to determine which entities to best collaborate with on this issue. The results of the running of the algorithms, described herein, a list of authoritative entities was generated. The second entity on this list was eventually utilized to solve the domain issue. The first entity on the list was not completely suited for the task due to security restriction.

The above presented scenario lends additional support towards the effectiveness of the authoritativeness metric. It is shown that the authoritativeness metric is applicable not only to academic, research, and scientific activities but also to integration of various domain expertise in a corporate setting as well. Further, it is shown that the list of authoritative entities is crucial for selection of collaborative partners due to various external constraints (e.g., security, geospatial reasons) that cannot be accounted for within the authoritativeness metric. By producing a list of authoritative entities the end user is capable of filtering for these external constraints while still achieving the results of finding the optimal collaborative partner.

7 Conclusions and Future Works

Open innovation is the means by which companies seek external entities with which to collaborate to form innovation. This paper illustrated that finding the best source of collaboration for a given innovation in a manual fashion is sub-optimal. This paper presented a novel methodology for the automation of collaboration partner detection for the purpose of collaborative innovation. Furthermore, a process by which the authoritativeness of the collaborative partner is ensured to be optimal was presented. Using data mining, clustering, and analysis of the documents related to the innovation domain increases competitiveness of companies.

Novel to this paper is the use of boosted authoritativeness. The iterative process of increasing, or decreasing the authoritativeness of possible candidates for collaborative innovation extends the search process, and represents an automated methodology for determining the best candidate entity (company, person) for collaborative innovation. Future research should includes increasing the efficiency of document detection during the web mining process as

well as increasing the rate at which document classification takes place.

Bibliography

[1] A. Kusiak, Innovation: A Data-Driven Approach, *International Journal of Production Economics*, Vol. 122, No. 1, pp. 440-448, 2009.

[2] G. Berkhout, P. van der Duin, Mobile Data Innovation: Lucio and the Cyclic Innovation Model, *Proc. of the 6th Intl. Conf. on Electorinc Commerce*, Delft, Netherlands, pp. 603-608, 2004.

[3] Y. Sawatani, F. Nakarmura, A. Sakakibara, M. Hoshi, S. Masuda, Innovation Patterns, *Proc. of the 2007 IEEE Intl. Conf. on Services and Computing*, Salt Lake City, UT, pp. 427-434, July 2007.

[4] J.B. Zhang, Y. Tao,The Interaction Based Innovation Process of Architectural Design Service, *Industrial Engineering and Engineering Management 2007 IEEE Intl. Conf.*, pp.1719 - 1723, Dec. 2007.

[5] A.W. Ulwick, Turn Customer Input Into Innovation. *Harvard Business Review*, Vol. 80, No. 1, pp. 91-97, 2002.

[6] L. Collins, Opening up the Innovation Process, *Engineering Management Journal*, Vol. 16, No. 1, pp. 14-17, 2006.

[7] A. Langville, C. Meyer, Deeper Inside PageRank, *Internet Mathematics*, Vol. 1, No. 3, pp. 335-380.

[8] B. Liu, *Web Data Mining*, Springer, Heidelberg, 2007.

[9] T. Haveliwala, Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 4, pp. 784-796, 2003.

[10] S. Kamvar, T. Haveliwala, C. Manning, G. Golub, Extrapolation Methods for Accelerating PageRank Computations, *Proc. Of the 12th Intl. Conf. on World Wide Web*, Budapest, Hungary, pp. 261-270, 2003.

[11] I. Whitten, E. Frank, *Data Mining, Practical Machine Learning Tools and Techniques*, Morgan Kauffman, New York, 2005.

[12] J. Han, Y. Yin, G. Dong, Efficient Mining of Partial Periodic Patterns in Time Series Databases, *Proc. of the 15th IEEE Intl. Conf. on Data Engineering*, Sydney, Australia, pp. 106-115, March 1999.

[13] S. Simon, Modeling and Design Aspects of the MEMS Switch, *Proc. Of the 2003 IEEE International Semiconductor Conference*, Sinaia, Romania, September 28 - October 2, pp. 128-132, 2003.

[14] R. Maeda, M. Takahashi, S. Sasaki, Commercialization of MEMS and Nano Manufacturing, *Proc. Of the 6th IEEE Intl. Conf. On Polymers and Adhesives in Microelectronics and Photonics*, Tokyo, Japan, pp. 20-23, January 2007.

[15] G. M. Rebeiz, Homepage $http://www.eecs.umich.edu/rebeiz/rebeiz.html$.

[16] C. L. Goldsmith, Homepage $http://www.memtronics.com/page.aspx?page_id=10$.

[17] Y. Zhai, B. Liu, Web Data Extraction Based on Partial Tree Alignment, *Proc. of the 2005 International World Wide Web Conference*, May 10-14. Chiba, Japan, pp. 76-85, 2005.

[18] L. Shih, D. Karger, Using URLs and Table Layout for Web Classification Tasks, *Proc. of WWW 2004*, May 17-22, New York, pp. 193-202, 2004.

[19] B. Kules, J. Kustanowitz, B. Shneiderman, Categorizing Web Search Results into Meaningful and Stable Categories Using Fast-Feature Techniques, *Proc. of JCDL'06*, June 11-15, pp. 210-219, 2006.

[20] X. Jin, R. Li, X. Shen, R. Bie, Automatic Web Pages Categorization with ReliefF and Hidden Nad've Bayes, *Proc. of SAC '07*, March 11-15, pp. 617-621, 2007.

[21] J. Engler, A. Kusiak, A. Mining the Requirements for Innovation, *Mechanical Engineering*, Vol. 130, No. 11, pp. 38-40, 2008.

[22] P. Chapman et al., *Step-by-step Data Mining Guide*, CRSIP-DM Consortium, CRISP-DM 1.0, 2000.

[23] N. Lavraĉ, H. Motoda, T. Fawcett, R. Holte, P. Langley, P. Adriaans, Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving, *Machine Learning*, Vol. 57, No. 1-2 , pp.13-34, 2004.

[24] R. Gajda, Utilizing Collaboration Theory to Evaluate Strategic Alliances, *American Journal of Evaluation*, Vol. 25, No. 1, pp. 65-77, 2004.

[25] M. Geringer, Strategic Determinants of Partner Selection Criteria in International Joint Ventures, *Journal of International Business Studies*, Vol. 22, No. 1, pp.755-786, 1991.

[26] M. Hitt, M. Dacin, E. Levitas, J. Arregle, A. Borza, Partner Selection in Emerging and Developed Market Contexts, *Academy of Management Journal*, Vol. 43, No. 3, pp. 440-467, 2000.

[27] Z. Chen, Learning about Learners: System Learning in Virtual Learning Environment, *International Journal of Computers, Communications and Control*, 3(1):33-40, 2008.

[28] H. Grebla, C. Cenan, C., Distributed Machine Learning in a Medical Domain, *International Journal of Computers, Communications & Control*, 1(S):245-250, 2006.