Use of the Normalized Word Vector Approach in Document Classification for an LKMC

Kevin R. Parker
Idaho State University
Pocatello, Idaho, USA

parkerkr@isu.edu

Robert Williams
Curtin University of Technology
Perth, WA, Australia

Bob.Williams@cbs.curtin.edu.au

Philip S. Nitse
Idaho State University
Pocatello, Idaho, USA

nitsphil@isu.edu

Albert S.M. Tay
Idaho State University
Pocatello, Idaho, USA

taysion@isu.edu

Abstract

In order to realize the objective of expanding library services to provide knowledge management support for small businesses, a series of requirements must be met. This particular phase of a larger research project focuses on one of the requirements: the need for a document classification system to rapidly determine the content of digital documents. Document classification techniques are examined to assess the available alternatives for realization of Library Knowledge Management Centers (LKMCs). After evaluating prominent techniques the authors opted to investigate a less well-known method, the Normalized Word Vector (NWV) approach, which has been used successfully in classifying highly unstructured documents, i.e., student essays. The authors propose utilizing the NWV approach for LKMC automatic document classification with the goal of developing a system whereby unfamiliar documents can be quickly classified into existing topic categories. This conceptual paper will outline an approach to test NWV's suitability in this area.

Keywords: Knowledge management, Competitive intelligence, Digital libraries, Document classification, Normalized Word Vector, Library as Knowledge Management Center, Small enterprises

Introduction

Many small businesses and entrepreneurs lack the time and resources to properly conduct competitive intelligence (CI) activities. Operational issues take up most of the owners' time and leave few other resources to devote to CI activities. There is, therefore, an opportunity for an outside entity to provide the needed services that will enable the small business to compete effectively. The concept of Libraries as Knowledge Management Centers (LKMC) was proposed to address problems faced by small busi-

nesses on one front, and by libraries on another (Parker, Nitse, & Flowers, 2005). At the core of the proposal is the premise that libraries can extend their services to act as knowledge management (KM) centers for small businesses, providing both KM and CI support. The arrangement would be beneficial both to libraries and to small businesses. Libraries benefit because it is an opportunity to reaffirm their relevance in a digital age in which so much information is freely available to patrons and library funding is deteriorating (ALA, 2004). Small businesses benefit because they are often unable to gather sufficient internal and external knowledge to assist in strategic planning and positioning, and thus are unable to compete with larger rivals whose resources allow them to develop sophisticated KM and CI systems. LKMCs hold promise to help level the playing field.

The seminal paper (Parker et al., 2005) enumerated the requirements that must be met for libraries to expand their services to act as KM centers for small businesses. This paper describes a single phase of the study, investigating the use of document classification techniques to classify and catalog digital documents for an LKMC. One of the linchpins of the LKMC is the ability to locate and retrieve pertinent information quickly. Therefore, accurate and efficient document categorization is an essential first step in the realization of an LKMC. The following section lays out the components of an LKMC, and explains each in detail.

Components of a Library Knowledge Management Center

As noted earlier, the seminal paper (Parker, Nitse, & Flowers, 2005) enumerated the requirements of an LKMC that must be met for the expansion of library services to include KM and CI offerings for small businesses. First, some businesses are associated with a particular jargon, and if such businesses are to be served by the LKMC then appropriate domain ontologies must be developed. Second, automatic document classification must be available to determine the content of both existing digital documents as well as new documents that are being delivered on a constant basis by streaming information sources. Next, library indexing or cataloging systems must be modified to incorporate conceptual details about documents so that Semantic Web technology can be used to semantically link the library's resources, making semantically related documents easier to retrieve and deliver. Each of these components will be briefly considered.

A domain ontology is a clearly stated formal specification of the basic concepts (objects, concepts, and relationships) that are known to exist in some area of interest. Specific domains can be identified and a common ontology can be defined to map vocabularies of specified terms with generally accepted definitions (Gruber, 1991). Tools like the Ontolingua Server are available to assist in the development of ontologies (Farquhar, Fikes, & Rice, 1997). Building a domain ontology for a specific business type requires a thorough understanding of the domain. Therefore the process should start by identifying general terms common to all small businesses, and then narrowing the focus to a specific business with the purpose of determining common industry terms, organization-specific terms, and even project-specific terms. A complete domain ontology spans a wide spectrum of corporate interests, thus providing the means to identify a greater percentage of relevant information. A specialist trained in knowledge engineering may be required to assist in the specification of key concepts for the domain ontology. Further, domain ontologies are already available for many industries.

Second, as digital documents are added to the library's collection for CI/KM purposes, document classification techniques can assist in determining the contents of each. The library's collection will consist of documents from a variety of external sources. These may include items stored at other library locations, or items provided by pay-for-use services such as Dow Jones, Hoover's Company Data Bank, Standards & Poor's, NewsEdge, or free information sources such as SEC's

Edgar system and corporateinformation.com (Breeding, 2000). The library may also subscribe to specialized databases from third-party vendors (Dialog, Lexus/Nexus) or press release and news-feed collections (WavePhore's Newscast Access or NewsEdge's NewsObjects), or offer access to product literature, competitor web sites, archived design specifications, company profiles and financial statements, and numerous other sources (Johnson, 1998). Internal information, such as internally generated knowledge "extracted" from the minds of the company's employees, must also be classified and stored. This type of information is typically not accessible by others and is often lost when an employee retires or leaves for other reasons. The LKMC can provide secure servers on which companies can store proprietary internal information in a structured and accessible fashion. An interface will be required to allow authorized company employees to store and access the proprietary knowledge.

Next, the library catalog system must be modified to store details about specific topics (concepts) and in what references to find them, because there may be many key topics or concepts in each reference. This may require significant changes because many libraries store only catalog details about what is in a particular reference. Semantic Web techniques will be used to semantically link the library's resources, so that semantically related documents can easily be retrieved or delivered.

Review of Document Classification Literature

Because the focus of this phase of the study is to investigate document classification techniques for use in an LKMC, a better understanding of document classification is required. Some of the earliest work in document classification took place in the early 1950s. In 1952, Luhn (1952) presented the first version of the "Luhn Scanner", also referred to as the IBM Electronic Information Searching System, and additional papers on the recording of and searching for literary information followed (Luhn, 1953, 1959). The automatic classification of documents is especially useful in the library environment. The early work of Maron (Maron, 1965; Maron & Kuhns, 1960), of Borko (Borko & Bernick, 1963, 1975, 1978), and later Larson (1992) and Plaunt and Norgard (1998), all attempt to automatically apply existing human-created ontologies, thesauri, and classification schemes to real library data. Studies in document classification have taken a variety of paths over the years. Several of the more common techniques will be described in the following section. The discussion is not exhaustive since new techniques like neural networks and Wikipedia-based approaches continue to be proposed.

Decision Trees

Decision trees consist of a set of rules that are applied in a sequential way until a decision is yielded (Hotho, Nürnberger, & Paaß, 2005). The training process starts with a comprehensive training set of labeled documents. A word is selected because it is deemed to best predict the correct document classification. The set is then partitioned into two subsets, the subset with documents containing the word, and subset containing the documents without the word. The procedure is recursively applied to each subset, stopping only when all documents in a subset belong to the same category.

Decision Rules

Decision rules, also referred to as symbolic rule learning, are the subject of many studies including (Apté, Damerau, & Weiss, 1994) and (Cohen & Singer, 1999). Brücher, Knolmayer, and Mittermayer (2002) explain that decision rule algorithms construct a rule set for every category. A single rule generally consists of a category name and a dictionary feature that is representative of the documents belonging to the category. Documents that satisfy a category's rules are assigned to that category.

Clustering Techniques

Jain, Murty, and Flynn (1999) provide an excellent explanation and comparison of clustering techniques, including hierarchical clustering, partitional algorithms, nearest neighbor clustering, fuzzy clustering, etc. The k-means approach (MacQueen, 1967) is the simplest and most commonly used partitional algorithm. The algorithm uses cosine similarity for document clustering. It starts with a random initial partition and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centers until a convergence criterion is satisfied. Nearest neighbor classification is usually performed by selecting documents from the training set that are "similar" to the target document. If k similar documents are considered, the approach is also known as k-nearest neighbor classification (Hotho et al., 2005).

Probabilistic Bayesian Models

Many approaches to document classification make use of statistical language modeling approaches like Bayesian classification techniques. Probabilistic classifiers are based on the assumption that the presence of words that make up a document is the result of a probabilistic mechanism, which means that the category into which a document falls has some relation to the words that appear in the document (Hotho et al., 2005). Bayesian classification uses training data to calculate Bayes optimal estimates of the model parameters, which are then used to classify new test documents by using Bayes rule to calculate the probability that a class would have generated the test document in question (Baker & McCallum, 1998).

Vector-Based Methods (Support Vector Machines)

In 1957 Luhn postulated that automatic text retrieval systems could be based on content identifiers attached to both the stored text and users' queries. With documents represented by term vectors, and queries by either term vectors or Boolean statements, a query-document similarity value can be obtained by assigning term weights and comparing the corresponding vectors. Much of Salton's work deals with similar vector space models, beginning in 1962 and continuing over the next 35 years (Salton, 1962; Salton, Singhal, Mitra, & Buckley, 1997). In one of the more widely cited papers, Salton and Buckley (1988) explain the importance of automatic term weighting and propose single-term-indexing models to which other content analysis procedures can be compared.

Dimensionality-Reduction Techniques

There are a variety of dimensionality-reduction techniques such as latent semantic indexing, or LSI (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). LSI is based on the implicit higher-order structure in the association of terms with documents, and works by applying matrix decomposition to a term-by-document matrix of the collection and generating a large number of orthogonal LSI factors. Associations among terms and documents are calculated with the assumption that there is an underlying structure in the pattern of word usage across documents (Schütze, Hull, & Pedersen, 1995).

Normalized Word Vector Approach

The NWV approach was developed in the early 2000s for use in the automatic grading of essays (Williams and Dreher, 2004). One aspect of essay grading involves the classification of unstructured documents, and the NWV technique has been tested extensively and successfully in essay classification and grading. The approach is discussed in detail in Williams (2006, 2007), Williams and Dreher (2004, 2005a, 2005b), and Dreher and Williams (2006), and an overview of the technique is presented in the following section.

The Normalized Word Vector Approach

All of the systems discussed above have their attractions, but recent studies have shown that there is little difference between the performances of the best text categorization systems, as if a plateau has been reached. Much of the work that continues on these systems achieves performance improvements of a few percentage points only (Gabrilovich & Markovitch, 2006). Therefore we decided to assess a different approach. The successful use of the NWV approach in the classification of unstructured documents for essay classification and grading made it worthy of consideration as a viable alternative for use in the LKMC.

A vector may be defined as a directed line-segment with a length and a direction (Smail, 1949). Vectors exist in a vector space, which can have 2 or more dimensions. A simple straight line in an XY graph is an example of a vector. Two vectors in the same space, both with their origins at (0, 0), form an angle. Measurement of the similarity of the vectors can be undertaken by taking the cosine of the angle.

The NWV concept can be explained as follows. Recall that it was stated earlier that vector-based methods represent documents by term vectors. NWVs are special vectors in a large dimensional space that represent the content of a document. The dimensions of the space are derived from the number of core concepts in a thesaurus – typically 800-1000 in a modern electronic thesaurus. In this case the dimension of the space is 812, representing the 812 concepts in the Macquarie Thesaurus (Macquarie, 2007).

The coordinates of each NWV can be computed by counting the number of times each of the 812 concepts occurs in the document under consideration. In order to determine the concept counts required to build this vector representation, each word in the document must be "normalized", or in other words reduced to a thesaurus root word appropriate to the concept. These concept counts are then used for the vector representation.

Once the vector representations have been calculated for each document they can be compared to the NWVs established for a set of standard or representative documents that are typical of the topics of interest in order to determine the degree of similarity for classification purposes. An important predictor of the similarity of document content is the cosine of the angle between the NWV representing the model document and the NWV representing the document being classified. For example, suppose we have a simple thesaurus with the following words assigned to one of 5 concept numbers:

Concept Number	Words
1.	the, a
2.	pretty, lovely, gorgeous
3.	flower, bloom, blossom
4.	red
5.	yellow

Next suppose we have the following successive sentence fragments from two separate documents:

Document#	Document Text	
(1)	The pretty flower…	A lovely bloom…
(2)	The red blossom…	A yellow bloom…

If we view the concept numbers as representing the axes of a five dimensional space, then the vectors for these two documents can be written by counting the number of times that a word associated with each concept number appears in the document fragments. In document 1 there are two words associated with concept 1 – 'the' in fragment 1 and 'a' in fragment 2. There are two words associated with concept 2 – 'pretty' in fragment 1, and 'lovely' in fragment 2. There are two words associated with concept 3 – 'flower' in fragment 1 and 'bloom' in fragment 2. In document 1 there are zero words in the fragments associated with concepts 4 and 5. Document 2 is assessed in a similar manner. The table below summarizes this analysis for both documents.

Document#	Vector Representation	Explanation
(1)	[2, 2, 2, 0, 0]	[the, a; pretty, lovely; flower, bloom; null; null]
(2)	[2, 0, 2, 1, 1]	[the, a; null; blossom, bloom; red; yellow]

Because graphical representations beyond three dimensions are difficult to produce the remainder of this discussion will consider only the first three dimensions for these documents. Thus, the first three dimensions give us:

Document#	Vector-first 3 concepts	Explanation
(1)	[2, 2, 2]	[the, a; pretty, lovely; flower, bloom]
(2)	[2, 0, 2]	[the, a; null; blossom, bloom]

The vectors for the first three dimensions are illustrated in Figure 1 as follows. Each axis represents one of the three concepts (or dimensions). Vectors 1 and 2 represent the documents and are instances of what are termed NWVs. Vector 1 represents a line from [0,0,0] through [2,2,2] and vector 2 a line from [0,0,0] through [2,0,2]. If we assume that document 1 is the exemplar document, then we can see how semantically close document 2 is to the exemplar document by looking at the closeness of their corresponding vectors. The angle between the exemplar document vector and the vector for document 2 is named theta, and its location is shown in Figure 1. The angle between the vectors varies according to how "close" the vectors are. A small angle indicates that the documents contain similar content; a large angle indicates that they do not share much common content.

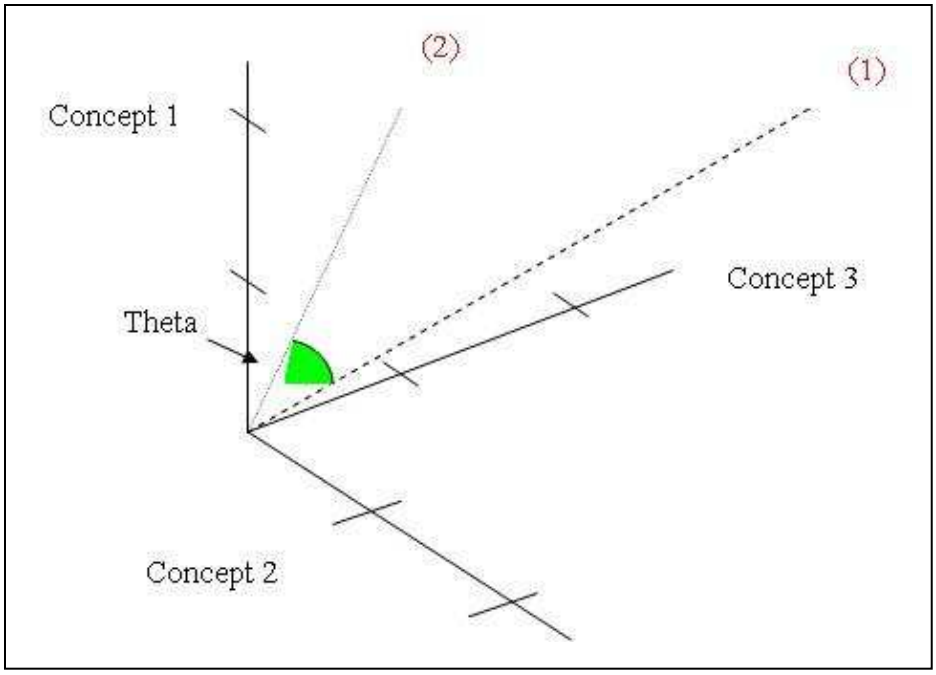

Figure 1. Vector representation (dashed lines) of documents

It turns out in practice that the cosine of theta is generally a powerful predictor of document similarity (Williams, 2006). If the two documents above were identical in terms of the number of times each concept was mentioned, then the NWVs would be identical and they would appear as collinear vectors in the diagram with a cosine equal to 1. If the documents were completely different, the vectors would be orthogonal, and their cosine would be 0. For these and all other cases the cosine of the angle is calculated in the normal fashion.

In general, these ideas are extended to the 812 concepts in the base thesaurus (Macquarie, 2007) and all words in the documents. Therefore, the vector space over which the document vectors are computed has 812 dimensions, and the vector theory carries over to these dimensions in exactly the same way – it is of course difficult to visualize the vectors in this hyperspace.

Future Testing of NWV for LKMC

The next step in this phase of the study will be to assess the performance of the NWV approach by conducting a case study of a small prototype implementation in a particular constrained domain. It will require selection of classification criteria, domain selection, training, selection of exemplar documents, and the classification itself. If the NVW proves to perform LKMC document classification quickly and accurately, the project will then move to another phase, studying how to best modify library indexing or cataloging systems to integrate conceptual document details so that Semantic Web technology can be used to semantically link the library's resources. If the NWV approach is not satisfactory, then other document classification techniques must be evaluated and tested.

Classification Criteria

The topics for classifying the documents are extracted from the Library of Congress Subject Headings (LCSH). The LCSH is a controlled vocabulary that provides subject access points to the bibliographic records contained in the Library of Congress catalogs. Only one heading represents each subject in the LCSH. Synonymous terms and variant forms of the same heading are included as entry vocabulary, i.e., as 'referred-from' terms. When an object, a concept, or a named entity is

known by more than one term (i.e., a word or phrase) or name, attempts are made to select for use as subject headings terms or names that are most likely to be sought by catalog users (The Library Corporation, 2007). In the current edition of the LCSH there are over 280,000 total headings and references (Library of Congress, 2007). Using the LCSH, documents for a selected topic will be easily identified. The headings and related terms from the LCSH will also be used as the initial vectors for each topic.

Domain Selection

Document classification will proceed as follows. A topic such as ethical behavior in operating a small business will be chosen to test the performance of the proposed system. The system will be trained with some exemplar documents, and then random documents will be analyzed for similarity/dissimilarity and assigned scores.

Training and Exemplar Documents

One hundred electronic versions of articles on a given topic will be selected from the web and other electronic sources. These will be read by humans, who will assign a score from 0 to 100 to each article. Higher scores will indicate articles with more relevant content for the topic; lower scores will indicate less topic content. These articles will then form the training set of articles.

A second set of 100 articles on the same topic will also be processed in the above manner. This set then forms the validation set of articles.

The training articles will each have a NWV computed. Several other features of each article, such as the number of words in the article, the number of adjectives, and the number of adverbs will also be measured.

Multiple linear regression will then be undertaken with the NWVs and the other features on the training set, with the score assigned by the humans as the dependent variable. The outcome of the regression analysis will be a scoring equation that will predict a score from the significant independent variables (features) for the given topic.

This equation will then be used with the relevant measures from the validation articles to predict their scores. These system-generated scores will then be compared to the human scores for the validation articles to check the accuracy of the derived equation. The correlation coefficient will be used to measure the usefulness of this predictor equation.

The scoring equation will then be stored for this topic. Scoring equations will also be built in the above manner for all topics to be considered for classification.

Classification

When a new article arrives for classification, its NWV will be calculated and predictor features will be measured. These will then be input into each of the stored scoring equations. This includes comparing the new document's NWV to the NWVs in the database by calculating the cosines between the document to be classified and the entries in the database. The article will then be assigned to the topic which produces, via its scoring equation, the highest score. We believe that this will be a unique and innovative method to classify documents and will especially attend to the documents' content and semantics.

Additional Features

We are also working on a feature whereby the user can expand or compress the basic thesaurus upon which the model is reliant. The user can construct a personal sub-domain of concepts from the thesaurus. The user may wish to find documents relating to particular concepts, and may wish

to define a particular set of semantic relationships among these concepts. The system will provide interactive support for this construction, and then search a target set of documents for these concepts, and highlight in these documents where these concepts occur. A report will then be generated to summarize the concept counts, the relative importance of the concepts as determined by the previously constructed concept hierarchy, and the proportion of the document related to these concepts.

The user will also be able to add new concepts to the thesaurus database. This will require the user to identify the new concepts, assign a concept number, provide synonyms for the concept, and classify the word entries as nouns, verbs, adjectives, and adverbs. In this way a specialized thesaurus can be constructed for a particular knowledge domain of interest to the user.

Conclusion

This conceptual paper describes the consideration of various document classification approaches to best implement an LKMC. To be successful, an LKMC must be capable of quickly responding to requests for information and of providing needed information in a timely manner. To accomplish this daunting task, LKMCs require appropriate tools to find, capture, and report the intelligence that is needed. The first step in this phase of the study is to select an appropriate document classification approach. Although there are many document classification techniques from which to choose, each has its own advantages and shortcomings, and therefore new approaches to document classification were considered. Since the NWV approach has been tested successfully in the classification and assessment of essays, it was deemed worth investigating to see if its success will carry over to document classification for the LKMC. Future research will test not only the document classification performance of the NWV approach, but also its extensibility into specialized domains. If the NWV approach proves to be suitable for an LKMC, then the concept of LKMCs will be closer to reality and the possibility that small businesses can compete on the basis of KM and CI resources will be greatly enhanced.

References

American Library Association. (2004). *ALA's report on library funding in the United States.* American Library Association, Chicago, IL.

Apté, C., Damerau, F., & Weiss, S. (1994a). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems, 12*(3), 233-251. Retrieved April 14, 2007 from http://www.research.ibm.com/dar/papers/pdf/tois94_with_cover.pdf

Baker, L. D., & McCallum, A. K. (1998). Distributional clustering of words for text classification. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 96-103. Retrieved April 14, 2007 from http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mccallum/www/papers/clustering-sigir98s.ps.gz

Borko, H., & Bernick, M. (1963). Automatic document classification. *Journal of the ACM, 10*(2), 151-162.

Borko, H., & Bernier, C. L. (1975). *Abstracting concepts and methods*. New York: Academic Press.

Borko, H., & Bernier, C. L. (1978). *Indexing concepts and methods.* New York: Academic Press.

Breeding, B. (2000). CI and KM convergence: A case study at Shell Services International. *Competitive Intelligence Review, 11*(4), 12-24.

Brücher, H., Knolmayer, G., & Mittermayer, M. (2002). Document classification methods for organizing explicit knowledge. *Proceedings of the Third European Conference on Organizational Knowledge, Learning and Capabilities*, Athens. Retrieved April 14, 2007 from http://www.ie.iwi.unibe.ch/staff/mittermayer/resource/Athen.pdf

Cohen, W.W., & Singer, Y. (1999). Context-sensitive learning methods for text categorization. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, 307-315. Retrieved April 14, 2007 from http://www.cs.cmu.edu/~wcohen/postscript/tois-sigir.pdf

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391-407. Retrieved April 14, 2007 from http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf

Dreher, H., & Williams, R. (2006). Assisted query formulation using normalised word vector and dynamic ontological filtering. In H. Legind Larsen et al. (Eds.), *FQAS 2006*, *LNAI 4027*, pp. 282-294. Berlin Heidelberg: Springer-Verlag.

Farquhar, A., Fikes, R., & Rice, J. (1997). The ontolingua server: A tool for collaborative ontology construction. *International Journal of Human-Computer Studies*, *46*(6), 707-728.

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, July 16–20, 2006, Boston, Massachusetts, 1301-1306. Retrieved May 30, 2007 from http://www.cs.technion.ac.il/~gabr/papers/wiki-aaai06.pdf

Gruber, T. (1991). The role of common ontology in achieving sharable, reusable knowledge bases. In J. A. Allen, R. Fikes, & E. Sandewall (Eds), *Principles of knowledge representation and reasoning* (pp. 601-602). San Mateo, CA: Morgan Kaufmann.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining, LDV Forum-GLDV. *Journal for Computational Linguistics and Language Technology*, *20*, 19-62. Retrieved April 14, 2007 from http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264-323.

Johnson, A. R. (1998). *An introduction to knowledge management as a framework for competitive intelligence.* Paper presented at the International Knowledge Management Executive Summit, San Diego, CA.

Larson, R. R. (1992). Experiment in automatic Library of Congress classification. *Journal of the American Society for Information Science*, *43*(2), 130-148.

Library of Congress. (2007). *Cataloging distribution service*. Retrieved November 20, 2007, from http://www.loc.gov/cds/lcsh.html

Luhn, H. P. (1952). *The IBM electronic information searching system.* IBM Research Center, Yorktown Heights, N.Y.

Luhn, H. P. (1953). A new method of recording and searching information. *American Documentation*, *4*(1), 14-16.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information, *IBM Journal of Research and Development*, *1*(4), 309-317. Retrieved April 14, 2007 from http://www.research.ibm.com/journal/rd/014/ibmrd0104D.pdf

Luhn, H. P. (1959). *Keyword-in-context index for technical literature (KWIC index).* Yorktown Heights, NY: IBM Advanced System Development Division, RC-127.

Macquarie Dictionary. (2007). Retrieved May 28, 2007 from http://www.macquariedictionary.com.au/

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley: University of California Press, 281-297.

Maron, M. E. (1965). Mechanized documentation: the logic behind a probabilistic interpretation. In M.E. Stevens, V. E., Giuliano, & L.B. Heilprin (Eds.), *Statistical association methods for mechanized documentation* (pp. 9-13). Washington, D.C.: National Bureau of Standards.

Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM, 7*(3), 216-244.

Parker, K. R., Nitse, P. S., & Flowers K. A. (2005). Libraries as knowledge management centers. *Library Management Journal -- Special Issue on Digital Libraries in the Knowledge Era: Knowledge Management and Semantic Web Technology, 26*(4/5), 176-189.

Plaunt, C., & Norgard, B. A. (1998). An association based method for automatic indexing with a controlled vocabulary. *Journal of the American Society for Information Science, 49*(10), 888-902. Retrieved from http://metadata.sims.berkeley.edu/assoc/assoc.ps

Salton, G. (1962). Manipulation of trees in information retrieval. *Communications of the ACM, 5*(2), 103-114. Retrieved April 14, 2007 from http://portal.acm.org/citation.cfm?id=366792.366828

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513-523. Retrieved April 14, 2007 from http://www.doc.ic.ac.uk/~jmag/classic/1988.Term-weighting%20approaches%20in%20automatic%20text%20retrieval.pdf

Salton, G., Singhal, A. Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing and Management, 33*(2), 193-207.

Schütze, H., Hull, D., & Pedersen, J. O. (1995). A comparison of classifiers and document representations for the routing problem. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95)*, 229-237. Retrieved April 14, 2007 from http://lsa.colorado.edu/~simon/LexicalSemantics/schutze95comparison.pdf

Smail, L. L. (1949). *Calculus*. New York: Appleton-Century-Crofts.

The Library Corporation. (2007). *Cataloger's reference shelf.* Retrieved November 20, 2007, from http://www.itsmarc.com/crs

Williams, R. (2006). The power of normalised word vectors for automatically grading essays. *Journal of Issues in Informing Science and Information Technology*, *3*, 721-729. Retrieved from http://informingscience.org/proceedings/InSITE2006/IISITWill155.pdf

Williams, R. (2007). A computational effective document semantic representation. *Proceedings of IEEE-Digital Ecosystems and Technologies 2007 Conference*, Cairns, Australia, 21-23 February.

Williams, R., & Dreher, H. (2004). Automatically grading essays with MarkIT *Journal of Issues in Informing Science and Information Technology*, *1*, 693-700. Retrieved from http://proceedings.informingscience.org/InSITE2004/092willi.pdf

Williams, R., & Dreher, H. (2005a). Formative assessment visual feedback in computer graded essays. *Journal of Issues in Informing Science and Information Technology*, *2*, 23-32. Retrieved from http://proceedings.informingscience.org/InSITE2005/I03f95Will.pdf

Williams, R. & Dreher, H. (2005b). Telecommunications use in education to provide interactive visual feedback on automatically graded essays. *Proceedings of International Telecommunications Society Africa-Asia-Australasia Regional Conference*, Perth, Australia.

Biography

Dr. Kevin R. Parker is a Professor of Computer Information Systems at Idaho State University, having previously held an academic appointment at Saint Louis University. He has taught both computer science and information systems courses over the course of his seventeen years in academia. Dr. Parker's research interests include competitive intelligence, knowledge management, the Semantic Web, and information assurance. He has published in such journals as *Journal of Information Technology Education, Journal of Information Systems Education*, and *Communications of the AIS*. Dr.

Parker's teaching interests include web development technologies, programming languages, and database management systems. Dr. Parker holds a B.A. in Computer Science from the University of Texas at Austin (1982), an M.S. in Computer Science from Texas Tech University (1991), and a Ph.D. in Management Information Systems from Texas Tech University (1995). Before entering academia Dr. Parker was a programmer/analyst with Conoco, Inc.

Robert Williams has over 25 years experience in the Information Systems industry as a practitioner, researcher and lecturer. He currently is a lecturer in the School of Information Systems at Curtin University of Technology in Perth, Western Australia. He has extensive experience in systems analysis and design and programming on a variety of mainframe, mini and personal computers, as well as a variety of operating systems and programming languages. He has published a number of articles on system users' personalities and satisfaction, decision support systems, and automated essay grading systems. Williams holds a BA degree with double majors in Mathematics and Economics from the University of Western Australia, a Graduate Diploma in Computing from the Western Australian Institute of Technology, and a Master of Information Systems degree from Curtin University of Technology.

Dr. Philip S. Nitse is a Professor and Chair of Marketing at the Idaho State University. His areas of research interest include competitive intelligence, knowledge management, healthcare marketing, and marketing management. He has been published in the *European Journal of Marketing*, *Competitive Intelligence Review*, *Marketing Intelligence and Planning*, *Journal of Health Care Marketing*, *Journal of Direct Marketing*, and *Advances in Marketing*. He has a BS in Marketing from Arizona State University, an MBA from Memphis State University, and a PhD in Marketing from The University of Memphis. In addition, he has over 18 years of sales and sales management experience with organizations such as Carrier Air Conditioning, Georgia Pacific, Mass Merchandisers, and VR Business Brokers.

Dr. Albert Tay is an Assistant Professor of Computer Information Systems at Idaho State University. His areas of research interest include Global Information Systems, Organizational Impact of IT, Software Development and Support, and Technology Adoption. He has a BS in Information Systems from Brigham Young University-Hawaii (1992), a MS in Decision and Information Systems from Arizona State University (1993), and a PhD in Communication and Information Sciences from University of Hawaii at Manoa (2006). Albert has over six years of IT management experience with organizations such as Adaptec Inc, Applied Materials Inc., and Koeneman Capital Management.