UNSTRUCTURED CONSTRUCTION DOCUMENT CLASSIFICATION MODEL THROUGH LATENT SEMANTIC ANALYSIS (LSA)

Tarek Mahfouz, Assistant Professor, tmahfouz@bsu.edu
Department of Technology, College of Applied Science & Technology, Ball State University, Muncie, Indiana, USA
Amr Kandil, Assistant Professor, akandil@purdue.edu
Division of Construction & Management, School of Civil Engineering, Purdue University, West Lafayette, Indiana, USA

ABSTRACT

The increasing sophistication and complexity of construction project mandates extensive coordination between different parties and produces massive amounts of documents in diversified formats. The efficient use of these documents has became inevitably needed. The first step in making these documents effectively usable, efficient classification methods need to be created. In an attempt to provide a robust document classification methodology for the construction industry, the current research developed automated construction document classifiers through Latent semantic Analysis (LSA). The analyses and models developed focused on two groups of construction textual documents. The first constitutes of documents with high variation in words like correspondences and meeting minutes. The second relates to documents of low word variations (standardized terminologies) like construction claims and legal documents. The present paper (1) examined the suitability of LSA algorithms; (2) developed truncated feature spaces for the utilized document sets; (3) developed a number LSA automated classification models; (4) developed C++ algorithm which processes the outputs of the developed classifiers to a class predictions; and (5) tested and validated the developed models. The developed models were validated through over all classification accuracy and were compared against Gold Standard of human agreement measures.

To that end, 16 LSA models were developed, out of which the four with the highest accuracy were chosen. These models attained relatively better results than previous researches in the surveyed literature. An overall accuracy of 91% and 87% were attained in the first and second groups of documents processed, respectively. The main finding of this paper represent a step in a line of research that targets developing a coherent and integrated methodology for Knowledge Management (KM) and construction decision support through Machine Learning (ML) techniques. It is conjectured that this research would help in relieving the negative consequences associated with lengthy processes of analyzing textual documents in the industry.

Keywords: Knowledge Management, Latent Semantic analysis, Machine Learning, Document Classification

1. INTRODUCTION

The US Census data showed that the total construction spending in 2007 was about $ 14 trillion (US Census 2010). This considerable amount of expenditure is constantly at risk due to the dynamic nature of the construction industry and the increasing sophistication and complexity of construction project. These two characteristics of the industry created a strong need for increasing the collaboration between diversified parties that may not exist in the same geographic region (Caldas et al. 2002). These characteristics also created a requirement for an extensive amount of coordination between the different parties, and the production of a massive amount of documents in diversified formats.

In an effort to mitigate the possible difficulties in managing projects with extensive documents, researchers have proposed artificial intelligence techniques for managing the knowledge these documents contain. Artificial Intelligence (AI) is being used to address increasingly sophisticated, diverse problems. It has been extensively

utilized to enhance information models, document integration, inter-organizational systems, and expert systems (Labidi 1997). A number of studies were undertaken using AI techniques to develop automated and semi-automated tools to enable the utilization of textual data expressed in natural language through text mining, document clustering, controlled vocabularies, and web-based models (Ioannou and Liu 1993, Yang et. al 1998, Caldas et. al 2002, Caldas and Soibelman 2003, and NG et al. 2006). Although those studies resulted in significant contribution, none of them investigated utilizing Latent Semantic Analysis (LSA) for the development of a generic model for unstructured construction document classification through automated extraction of novel knowledge.

Therefore, in an attempt to provide a robust document classification methodology for the construction industry, this paper developed automated classifiers through Latent semantic Analysis (LSA). The analyses and models developed in this paper focused on two groups of construction documents. The first constitutes of documents with high variation in words like transmittals, correspondences, and meeting minutes. The second group relates to documents of low word variations like construction claims and legal documents. This paper (1) investigated Latent Semantic Analysis (LSA) algorithms; (2) developed truncated feature spaces for the utilized document sets; (3) developed a number LSA automated classification models; (4) developed C++ algorithm which processes the outputs of the developed classifiers to a class predictions; and (5) tested and validated the developed models. It is conjectured that this research stream would help in relieving the negative consequences associated with lengthy process of analyzing textual documents in the construction industry.

2. BACHGROUND

2.1 Literature Review

Most available construction information integration tools are designed to work with structured data like CAD models and scheduling databases. However, a lot of important information is contained in semi-structured or unstructured format like contract documents, change orders, and meeting minutes, all of which are normally stored as text files (Caldas et al. 2002, Caldas and Soibelman 2003, and Al Qady and Kandil 2009). Consequently, facilitating the use of these documents through integrated methods has become a necessity to enhance project control, performance, and data reuse. A number of research studies have addressed this issue. Ioannou and Liu (1993) proposed a computerized database for classifying, documenting, storing and retrieving documents on rising construction technologies. Kosovac et al. (2000) investigated the use of controlled vocabularies for the representation of unstructured data. Wood (2000) provided a method for the hierarchical structuring of concepts extracted from textual design documents. Scherer and Reul (2002) utilized text mining techniques to classify structured project documents. Caldas et al. (2002) and Caldas and Soibelman (2003) used information retrieval via text mining techniques to facilitate information management and permit knowledge discovery through automated categorization of various construction documents according to their associated project component. Xie et al. (2003) provided an integrated model for retrieving construction project documents to facilitate decision-making, logical judgment, and control for project managers. Caldas et al. (2005) proposed a methodology for incorporating construction project documents into project management information systems using semi-automated support integration to improve overall project control. To facilitate and improve design reuse, Demian and Fruchter (2005) investigated the use of different text analysis methodologies to highlight and quantify potential similarities among objects from an archive of building models. Ng et al. (2006) implemented Knowledge Discovery in Databases (KDD) through a text mining algorithm to define the relationships between type and location of different university facilities, and the nature of the required maintenance reported in the Facility Condition Assessment database. Zhu et al. (2007) employed text analysis and statistical techniques to improve information processing in construction projects by capturing key concepts and their patterns of occurrence within unstructured documents such as contract documents and provisions, and developed a metadata model to present associations between requests for information documents (RFIs).

It is clear from the above that text-mining techniques have a promising potential in automating the handling of construction documents. However, none of these researches explored the potential of LSA. Consequently, the

goal of this research is to develop automated classification models for the construction industry with a higher accuracy by utilizing LSA, an AI technique based on mathematical modeling and machine learning.

2.2 Latent Semantic Analysis (LSA)

"Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the meaning of words" (Landauer et al 2007). It attempts to model the mechanism of exactly how words and passage meanings can be constructed from experience with language. A corpus of related text imposes constraints on the meaning and semantic similarities of a word. For example, a word like "bank" can mean "a river side" or "an institution for financial transactions" based on the constraints imposed by the rest of words within a body of text. The theory of LSA hypothesizes that the meaning of a text is conveyed by the words from which it is composed. Therefore, it is based on determining the meaning of a word by solving these constraints in a mathematical form by utilizing linear algebra, particularly, singular value decomposition (SVD).

LSA is based on the concept of Vector Space Model implemented by Support Vector Machines. However, the main advantage in LSA is that it utilizes a truncated space in which the number of features is decreased. LSA represents word and passage meanings in a form of mathematical averages. Word meanings are formulated as average of the meaning of all the passages in which it appears, and the meaning of a passage as average of the meaning of all the words it contains. LSA methodology applies SVD for the reduction of dimensionality in which all of the local word context relations are simultaneously represented. LSA, unlike many other methods, employs a preprocessing step in which the overall distribution of a word over its usage contexts, is first taken into account independent of its correlations with other words.

LSA then implements three well defined steps. Firstly, text document within a training corpus are represented in a form of matrix (Figure 1).

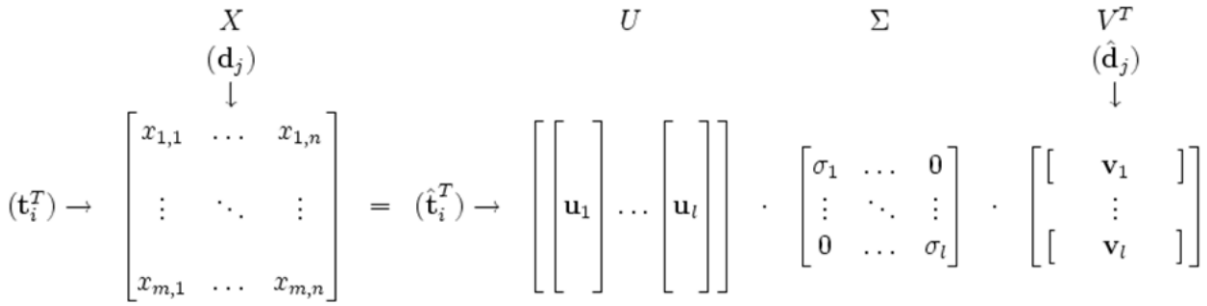
$$\mathbf{t}_i^T \rightarrow \begin{bmatrix} & \mathbf{d}_j & \\ & \downarrow & \\ x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} \quad \mathbf{t}_i^T = \begin{bmatrix} x_{i,1} & \cdots & x_{i,n} \end{bmatrix} \quad \mathbf{d}_j = \begin{bmatrix} x_{1,j} \\ \vdots \\ x_{m,j} \end{bmatrix}$$

Where:
The dot product between two term vectors $t_i^T t_p$ gives the correlation between the terms over the documents Input.
The dot product between two sentence vectors $d_j^T d_q$ gives the correlation over the terms Input.

Figure 1: Matrix representation in LSA (Landauer et al. 2007)

Each row of the developed matrix demonstrates a specific word in the training corpus. Each column of the matrix stands for a text document. Each cell contains the frequency with which the word of its row appears in the passage denoted by its column (Landauer et al. 2007). Often, the number of features m is much higher than the number of documents n within the collection. Removal of stop words before performing matrix representation is not a necessity, due to the mathematical nature of the SVD, but it enhances its performance by removing excess noise. The developed m by n matrix will contain zero and nonzero elements. Generally, a weighing function is applied to nonzero element to give lower weights to high frequency features that occur in many documents and higher weights to features that occur in some documents but not all (Salton and Buckley, 1991). Weighing functions are of two types namely local and global. The former relates to increasing or decreasing a nonzero element with respect to each document. The latter relates to increasing or decreasing a nonzero element across the whole collection of documents.

Secondly, SVD is applied to the developed matrix to achieve an equivalent representation in a smaller dimension space (Choi et al. 2001). With SVD, a rectangular matrix is decomposed into the product of three other matrices (Figure 2).



Where:

- The term-document matrix *X* represents how important a term is in a given document.
- Σ is a diagonal matrix representing the weights of the concepts. Usually SVD algorithm produces a Σ with $\sigma_1>\sigma_2>...>\sigma_k > 0$.
- The columns of *U* are the terms in concept space, the rows of *V* are the documents in the concept space. Since SVD is used, *U* and *V* are unitary matrices, that is, the rows of *U* and the columns of *V* are of unit length and are pairwise orthogonal.
- Not all concepts are necessarily used. Only those with sufficiently large singluar values (i.e. σ-s) may be selected. The result matrix is the best low-rank approximation to *X* in Frobenius norm.

Figure 2: SVD Matrix representation in LSA (Dumais 1990)

One component matrix describes the original row entities as vectors of derived orthogonal factor values, another describes the original column entities in the same way, and the third is a diagonal matrix containing scaling values such that when the three components are matrix-multiplied, the original matrix is reconstructed (Hofmann 1999). Thirdly, the number of features adopted for analysis is determined (Truncation). Since the singular value matrix is organized in an ascending order based on the weight of each term, it is easy to decide on a threshold singular value below which terms significance is negligible, refer to (Figures 2 and 3), (Dumais 1990). For an original matrix **A** with rank **k**, a newly truncated matrix $\mathbf{A_k}$ can be formulated by the dot product illustrated in equation 1.

$$A_i=\sum_{i=1}^{k} u_i\sigma_i v_i^T \rightarrow A_k=U_k\Sigma_k V_k^T \qquad (1)$$

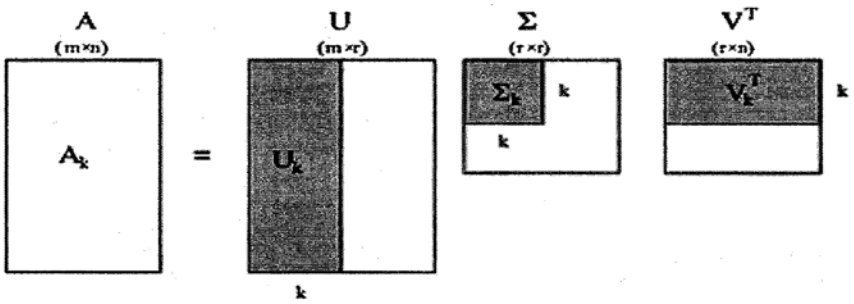


Figure 3: K Dimensional space representation in LSA (Dumais 1990)

By representing any document in the generated concept space, it is then possible to calculate "distance" on the set of such document representations thus computing whether two such representations are close which usually implies that the documents themselves are related. This notion makes LSA a very strong tool for document

classification. Landauer et al. (2007) highlights the superiority of LSA over other ML techniques with respect knowledge simulation. LSA has shown to reflect human knowledge in a variety ways (1) its measures highly correlate to human scores on standard vocabulary and subject matter tests; (2) it resembles human word sorting and category judgment; and (3) it accurately estimates passage coherence. LSA has been extensively used in linguistic researches. Landauer et al. (2003a and 2003b) tested LSA in multiple-choice vocabulary tests and the task of determining the adequacy of expository essays contents. In other studies, LSA successfully modeled several laboratory findings in cognitive psychology (Howard et al 2008).

3. METHODOLOGY

The adopted research methodology under the current task is composed of three main stages (Figure 4): (1) LSA Feature Space Development; (2) Model Design and Implementation; and (3) Model Testing and Validation.

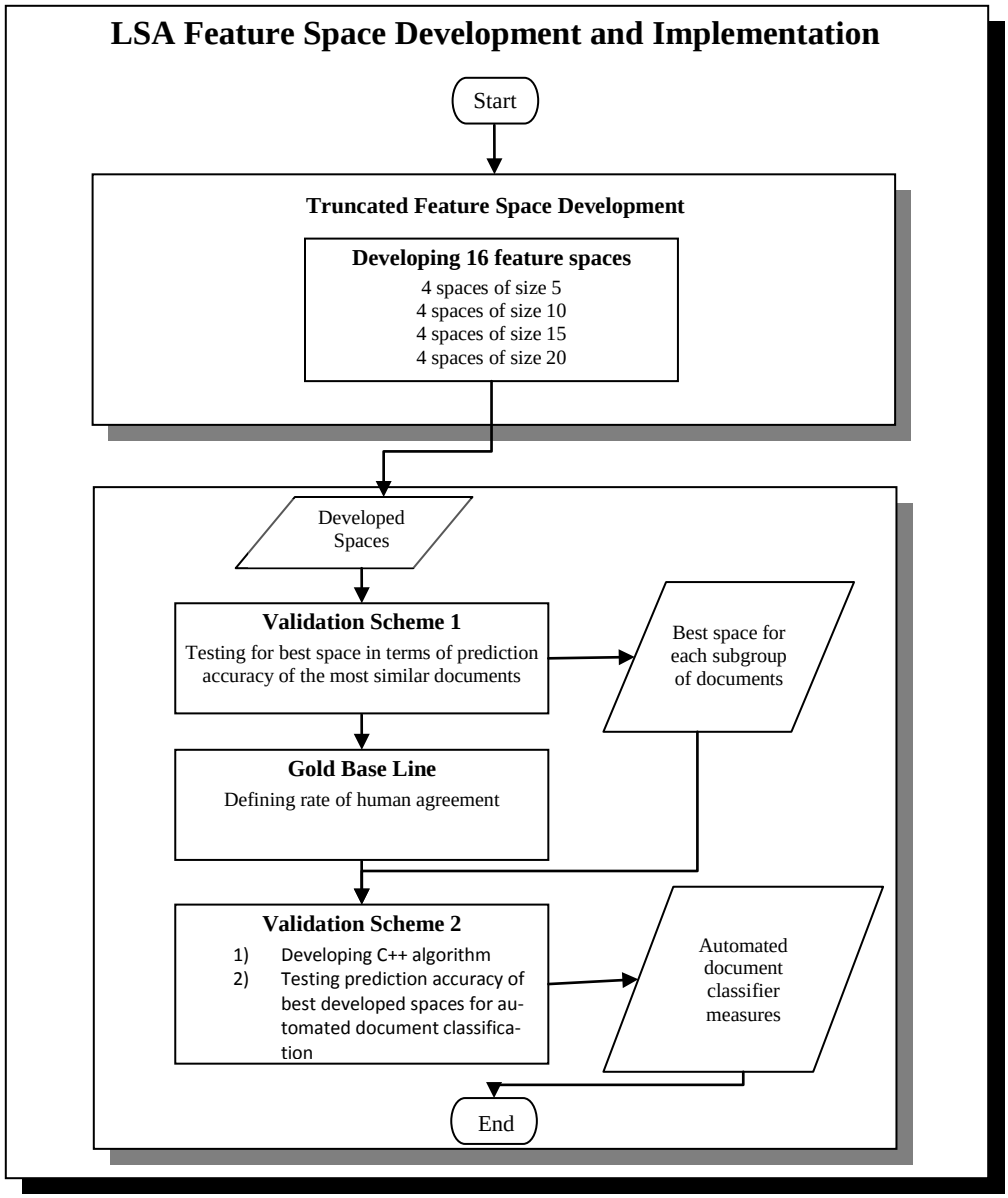


Figure 4: Research methodology

3.1. LSA Feature Space Development

In this research task an important parameter for the LSA model implementation is evaluated. Feature space in LSA is defined by the number of features that are used to represent a document as a vector. Under the current research task, features constitute of words within a document. Researches concerned with LSA feature space development cover wide range of truncated feature space sizes that enhance the effectiveness of the algorithm. It was highlighted in the literature that for dispersed dataset, a large feature space sizes ranging between 100 and 500 are appropriate (Choi et al. 2001). However, for closely related dataset, a feature space as small as 7 is appropriate (Koll 1979). In an earlier research performed by the authors, it was found that a feature space of size 10 is the best to represent a corpus of Differing Site Conditions (DCS) cases (Mahfouz 2009). Consequently, the present research task developed 4 different truncated feature spaces utilizing 5, 10, 15, and 20 features. As mentioned earlier, the current research task is concerned with two types of unstructured documents of high and low word variations. To that end, the first group constituted of 2 subgroups, the first of which included 300 correspondences and the second constituted of 150 meeting minutes. Furthermore, the second group constituted of 25 claims and 300 DSC cases as the first and second subgroups respectively. The four feature spaces were developed and tested for each subgroup of documents yielding 16 truncated feature spaces.

3.2. Model Design and Implementation

The following is a description of the steps of the LSA algorithm implemented for development of the automated document classifier. The algorithm starts with an argument, filename, which is the name of the file or directory to be parsed. The algorithm moves sequentially through each document, extracting relevant features (words), and excluding irrelevant ones which are included in a predefined list of words and characters. It converts all letters to lower case. It is worth mentioning that features that are not included in the common word list are considered to be relevant only if they are comprised of more than 2 characters. In addition, features of more than 20 characters are truncated to a maximum size of 20 characters.

After extracting relevant features and associating each one with the document it was extracted from, the algorithm begins calculating term weights. The (global) weights of the terms are computed over the collection of documents. By default, only a local weight is assigned and this is simply the frequency with which the term appears in a document. The algorithm implements two thresholds for term frequencies: Global and Local (GTP 2008). The implementation parameters of the algorithm are defined so that the global and local thresholds are both 2. A term must appear more than 2 times in the entire collection and in more than 2 document in the collection before it will be weighted. Next, the local weights of the features are computed. Each word weight is the product of a local weight times a global weight. Next, the algorithm creates a term-by-document matrix using the Harwell-Boeing sparse matrix format. The algorithm finally performs SVD decomposition.

Sixteen truncated feature spaces were generated. Each truncated feature space was generated with a local threshold of Log function and a global threshold of Entropy function. The Log function (equation 2) decreases the effect of large differences in term frequencies (Landauer et al 2007). The entropy function (equation 3), on the other hand, assigns lower weights to words repeated frequently over the entire document collection, as well as taking into consideration the distribution of each word frequency over the documents (Landauer et al. 2007). These thresholds were adopted for the current analysis due to their success over other types of threshold combinations and in earlier researches performed by the authors (Mahfouz 2009). Dumais (1991) illustrated that the log-entropy threshold combination attained 40% higher retrieval precision over other threshold combinations.

$$ltf_{i,d} = 1 + \log\left(tf_{i,d}\right);\ tf_{i,d} > 0 \tag{2}$$

$$\sum_i \frac{P_{ij}\log_2(P_{ij})}{\log_2 n} \text{ where } P_{ij} = \frac{tf_{ij}}{gf_i} \tag{3}$$

Where tf_{ij} is the word frequency of word i in document j, and gf_i is the total number of times that the word i appears in the entire collection of n documents.

General Text Parser (GTP) windows version, developed by Stefen Howard, Haibin Tang, Dian Martin, Justin Giles, Kevin Heinrich, Barry Britt, and Michael W. Berry, was utilized for the implementation of LSA feature spaces development. GTP is a general purpose text parser with matrix decomposition option which can be used for generating vector space information retrieval models. As stated by Landauer et al. (2007) "GTP could be considered the reference program for LSA analysis because it is a rewrite of the older Telcordia suite in more modern way. It is a very large program. Contrary to what its name indicates, GTP is not only a parser: It actually can run an SVD at the end of the process. GTP is a 100% C++ code".

3.3. Model Testing and Validation

The developed feature spaces were tested and validated in two folds. The first form of testing evaluated the LSA algorithm's ability to extract one of the documents originally used in the development of the feature space and the most related 2 documents from the data set. The second form of evaluation is based on correctly predicting the subject matter of newly untested documents from each subgroup. A C++ algorithm is developed to perform this step. The algorithm implemented by this program works in the following manner.

a) Each document in the feature space is tagged with a subject matter. The algorithm iterates sequentially through the documents storing the document number and its corresponding subject matter.
b) The LSA algorithm is implemented to extract the closest set of documents to the newly tested one. A prediction accuracy threshold of 97% is considered. In other words, any document retrieved at a similarity measure of less than 0.95 is disregarded. The algorithm is set to retrieve a document number and similarity measure (prediction accuracy %).
c) The program reads through the document number attained from the LSA implementation and retrieves the subject matter of each document.
d) The program reports the subject matter of the newly tested documents by two means. The first is reported as the most repeated subject matter. The second is reported as subject matters and weights, which are calculated based on the frequency of repetition of each subject among the retrieved documents. The reported outputs are compared against manual tagging of the newly tested document to decide on the most accurate method.

4. RESULTS AND DISCUSSION

The outcomes of the implementation of the aforementioned methodology are illustrated in tables 1 and 2.

4.1. Scheme 1 for Validation and Testing

In regards to the first test of validation, a closer examination of table 1 illustrates the followings.

a) With respect to correspondences similarity measures, a truncated feature space of 20 features attained 4%, 3%, and 1% enhancement of prediction accuracy over 5, 10, and 15 feature spaces respectively.
b) With respect to meeting minutes similarity measures, a truncated feature space of 20 features attained 7%, 4%, and 2% enhancement of prediction accuracy over 5, 10, and 15 feature spaces respectively.
c) With respect to claims similarity measures, a truncated feature space of 10 features attained 1%, 1%, and 2% enhancement of prediction accuracy over 5, 15, and 20 feature spaces respectively.
d) With respect to claims similarity measures, a truncated feature space of 10 features attained 1%, and 2% enhancement of prediction accuracy over 15, and 20 feature spaces respectively. There was no noticed enhancement in comparison to the 5 feature space.

Table 1: Prediction accuracy Measures of the First Scheme for Testing and Validation

Truncated Feature Space size	Average Prediction Accuracy %			
	Group 1		Group 2	
	Correspondences	Meeting minutes	Claims	DSC Cases
5	96	93	99	100
10	97	96	100	100
15	99	98	99	99
20	100	100	98	98

The attained results are attributed to the fact that the first group of analyzed document are comprised of highly variable terms. Correspondences and meeting minutes do not follow well standardized set of concepts like legal terms. On the other hand, they are comprised of natural language representations of human thoughts, meanings and intentions that are represented in the form of words. In addition, meeting minutes usually address a variety of topics and issues, which increases the complexity of the analysis. Consequently, a larger truncated feature space is required to capture the variability in the linguistic representations. On the other hand, the second group of analyzed documents constitute of words that follow structured and standardized format like legal and formulated engineering terms. Such aspect decreases the complexity of the analysis and yields better prediction accuracy at a lower truncated feature space.

4.2. Scheme 2 for Validation and Testing

The first step under this subtask established a Golden Standard of human agreement to which the performance of the developed model is to be compared. To that end, a set of 8 volunteers comprised of Assistant Professors, graduate students, and undergraduate students in construction engineering and management programs were utilized to set the base level of human agreement. It could observer that by virtue of the occupations of the participating volunteers, they posses enough knowledge about construction practices and documents. Each volunteer was provided a set of documents from each subgroup and asked to classify them according to similarities under related topics of his\her determination. A documents is considered to be classified correctly under a specific topic if three or more persons agreed on the document's topic (Alqady and Kandil 2009). The average agreements between participating members in regards to each set of documents are illustrated in the third columns of table 2.

Table 2: Golden Baseline of Human Agreement

Document Type		Average Agreement Between Humans	Average Prediction Accuracy of Developed Models	Truncated Feature Space Size
Group 1	Correspondences	97%	91%	20
	Meeting Minutes	89%	80%	20
Group 2	Claims	91%	85%	10
	DSC Cases	94%	87%	10

It is manifestly clear from table two that the lowest agreements were attained in relation to Meeting Minutes and Claims. Such aspect is attributed to that fact that these documents are usually comprised of a set of aspects that could not be defined under a specific title. A construction claim for example might include different causes of disputes that might not be related in nature.

The second step under this subtask tested the prediction accuracy of the developed truncated feature spaces, with the highest prediction accuracy achieved in section 4.1, for each set of documents as per the aforementioned methodology in section 3.3. The average accuracy of the developed models are illustrated in the fourth columns of table 2 and figure 5. A closer look at the attained results illustrate the followings.

a) The highest precision accuracy of the developed truncated feature spaces is 91% attained with respect to correspondences classification.
b) The lowest precision accuracy of the developed truncated feature spaces is 87% attained with respect to meeting minutes classification.
c) The prediction accuracy of the developed feature spaces were 6% to 9% lower than the human agreement levels (refer to figure 5)

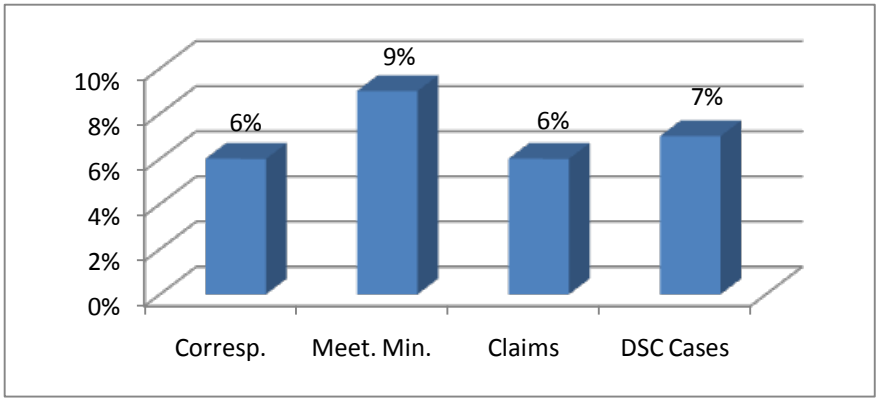


Figure 5: Variation Between Prediction Models and Human Agreements

It could be deduced from the above results that the developed model were consistent with the human prediction. Both attained the highest prediction and agreements in regards to correspondences and the lowest in regards to meeting minutes. Such aspect supports the arguments discussed in section 4.1 about the complexity of the classification problem under investigation.

5. SUMMARY AND CONCLUSION

The objective of this paper was to present an automated construction document classifier using Latent Semantic Analysis (LSA). To that end, a model was developed for each type of the four tested set of construction documents. the implementation of the aforementioned methodology highlights the feasibility and strength of LSA for automated document classification and extraction of latent knowledge from textual representation. In ongoing and future research efforts, the authors wish to expand the scope of the present models to cover analysis of time and money savings that could be achieved by the implementation of the aforementioned methodology.

REFERENCES

Al Qady, M. and Kandil, A. (2009). "Techniques for evaluating automated knowledge acquisition from contract documents." *In Proceedings of the Construction Research Congress*, ASCE, Reston, Va., 1479-1488.

Caldas, C. H., Soibelman, L., and Han, J. (2002). "Automated classification of construction project documents." *Journal of Computing in Civil Engineering*, 16(4), 234-243.

Caldas, C. H., and Soibelman, L. (2003). "Automating hierarchical document classification for construction management information systems." *Automation in Construction*, 12(4), 395-406.

Caldas, C. H., Soibelman, L., and Gasser, L. (2005). "Methodology for the integration of project documents in model-based information systems." *Journal of Computing in Civil Engineering*, 19(1), 25-33.

Choi, F. Y. Y., Wiemer-Hastings, P., and Moore, J. (2001). "Latent semantic analysis for text segmentation." *In Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, 109–117.

Demian, P., and Fruchter, R. (2005). "Measuring relevance in support of design reuse from archives of building product models." *Journal of Computing in Civil Engineering*, 19(2), 119-136.

Dumais, S. (1990). "Improving the information retrieval from external sources." *Behavior Research Methods and Computers*, 23, 229-236.

Dumais, S. (1991). "Improving the retrieval of information from external sources." *Behavior Research Methods, Instruments, and Computers*, 23(2), 229-236.

GTP. < http://www.cs.utk.edu/~lsi/soft.html> (Accessed 2008).

Hofmann, T. (1999). "Probabilistic latent semantic indexing." *Proceedings of the National Academy of Science*, 101, 5228-5235.

Howard, T. J., Culley, S. J., and Dekoninck, E. (2008). "Describing the creative design process by the integration of engineering design and cognitive psychology literature." *Journal of Design Studies*, 29(2), 160-180.

Ioannou, P. G., and Liu, L. Y. (1993). "Advanced construction technology system—ACTS." *Journal of Construction Engineering and Management*, 119(2), 288-306.

Koll, M. (1979). "An approach to concept-based information retrieval." *ACM SIGIR Forum, XIII*, 32-50.

Kosovac, B., Froese, T., and Vanier, D. (2000). "Integrating heterogeneous data representations in model-based AEC/FM systems." *Proceedings CIT 2000*, Reykjavik, Iceland, 1, 556-566.

Labidi, S. (1997). "Managing multi-expertise design of effective cooperative knowledge-based system." *Proc., 1997 IEEE Knowledge & Data Engineering Exchange Workshop*, IEEE, Piscataway, NJ, 10-18.

Landauer, T. K., Laham, d., and Foltz, P. W. (2003a). "Automated Essay Assessment." *Assessment in Education: Principles, Policy and Practice*, 10(30, 295-308.

Landauer T. K., Laham, d., and Foltz, P. W. (2003b). "Automated scoring and annotation of essays with the intelligent essay assessor." *Automated Essay Scoring: A Cross-disciplinary Prospective*, Shermis, M. D., and Burstein, J., editors, Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W. (2007). "Handbook of latent semantic analysis." *Lawrence Erlbaum Associates, London*.

Mahfouz, T. (2009). "Construction legal support for differing site conditions (DSC) through statistical modeling and machine learning (ML)"*Ph. D. thesis*, Department of Civil, Construction, and Environmental Engineering, *Iowa State Univ., Ames, IA.*

Ng, H. S., Toukourou, A., and Soibelman, L. (2006). "Knowledge discovery in a facility condition assessment database using text clustering." *Journal of Computing in Civil Engineering*, 12(1), 50-59

Salton, G., and Buckley, C. (1991). "Automatic text structuring and retrieval – experiment in automatic encyclopedia searching." *Proceeding of the 14th Annual International ACM SIGIR Conference on Research and Development in information Retrieval*, 21-30.

Scherer, R. J., and Reul, S. (2000). "Retrieval of project knowledge from heterogeneous AEC documents." *Proceedings of the Eight International Conference on Computer in Civil and Building Engineering*, Palo Alto, Calif., 812-819.

US Census Bureau < http://www.census.gov/const/www/c30index.html> (Accessed 2010)

Wood, W. H. (2000). "The development of modes in textual design data." *Proc., Eight International Conference on Computer in Civil and Building Engineering*, Palo Alto, Calif., 882-889.

Xie, H., Isaa, R. A., and O'Brien W. (2003). "User model and configurable visitor for construction project information retrieval." *4th Joint International Symposium on Information Technology in Civil Engineering*, ASCE, Nashville, Tennessee, 47.

Yang, M. C., Wood, W. H., and Cutkosky, M. R. (1998). "Data mining for thesaurus generation in informal design information retrieval" *Proceedings of the International Computing Congress*, ASCE, Reston, Va., 189-200.

Zhu, Y., Mao, W., and Ahmad, I. (2007). "Capturing implicit structures in unstructured content of construction documents." *Journal of Computing in Civil Engineering*, 21(3), 220-227.