Comparison of Latent Semantic Analysis and Vector Space Model for Automatic Identification of Competent Reviewers to Evaluate Papers

Yordan Kalmukov
Department of Computer Systems and Technologies
University of Ruse
Ruse, Bulgaria

***Abstract*—The assignment of reviewers to papers is one of the most important and challenging tasks in organizing scientific events. A major part of it is the correct identification of proper reviewers. This article presents a series of experiments aiming to test whether the latent semantic analysis (LSA) could be reliably used to identify competent reviewers to evaluate submitted papers. It also compares the performance of the LSA, the vector space model (VSM) and the method of explicit document description by a taxonomy of keywords, in computing accurate similarity factors between papers and reviewers. All the three methods share the same input datasets, taken from real-life conferences and the produced paper-reviewer similarities are evaluated with the same evaluation methods, allowing a fair and objective comparison between them. Experimental results show that in most cases LSA outperforms VSM and could even slightly outperform the explicit document description by a taxonomy of keywords, if the term-document matrix is composed of TF-IDF values, rather than the raw number of term occurrences.**

Keywords*—*Latent semantic analysis; vector space model; automatic assignment of reviewers to papers

I. Introduction

The assignment of reviewers to papers is probably the most important and challenging task in organizing the review process of scientific publications. Its accuracy has a direct impact on the conference/journal's quality and reputation. Submitted papers should be fairly evaluated by the most competent, in their subject domains, reviewers. To achieve that, the Program Committee (PC) chair or the assignment algorithm needs to know precisely the areas of expertise of all reviewers and the subject domains of all submitted papers. If the number of papers and reviewers is low, and all participants belong to some professional community, then it seems possible that the PC chair knows everybody, their areas of research and assigns reviewers to papers manually. However, when the number of papers and reviewers get higher, the manual assignment becomes highly inaccurate due to the lack of enough a-priori information and the many constraints (expertise, load-balancing, conflict of interests and etc.) that should be taken into account. In that case, the automatic assignment is the only accurate option. Its accuracy depends on both the assignment algorithm and the method of describing papers and reviewers' competencies. Assignment algorithms are studied in details in [1] and will not be discussed here. Instead, this article focuses on the methods of describing papers and identifying reviewers' competencies. Yes, both terms, describe and identify, are usable since methods could be explicit (users explicitly describe their papers or competencies) and implicit (subject domains and competencies are automatically identified by some piece of software).

Explicit methods usually rely on selection of keywords from a predefined list or a taxonomy [2]. They do not suffer from lack of information or sparse information, but could be a subject of incorrect, or even intentionally misleading, self-classification. Generally, choosing keywords from a predefined taxonomy of topics provides quite accurate calculation of paper-reviewer similarity factors [2].

In contrast, implicit methods do not require any additional description or actions from authors and reviewers. Instead, they rely on content analysis of both the submitted papers and the reviewers' previous publications. Implicit methods were somewhat inapplicable in the past, because reviewers whose publications cannot be found on the Internet will get their papers assigned to them at random. Currently this is not an issue anymore since all papers are published online and (at least) their abstracts are freely accessible. Fortunately, there are data aggregators such as Google Scholar, DBLP and Semantic Scholar. The latter provides an API that allows easy access to all abstracts of papers, published by a specified scientist, searching by name.

The aim of this paper is to experimentally test whether the latent semantic analysis (LSA), also known as latent semantic indexing (LSI), could be used for automatic identification of reviewers, competent to evaluate specific papers, and compare the results (in terms of accuracy) to the ones of the much simpler vector space model (VSM). The analyses are performed over real datasets taken from the CompSysTech series of conferences for a period of 5 years - from 2014 to 2018.

The paper is organized as follows: Section 2 discusses previous work from other researchers. Section 3 provides some details of how the vector space model could be used to identify competent reviewers to evaluate papers. Section 4 gives similar information but related to the use of latent semantic analysis for identifying reviewers. Section 5 describes the experimental setup and Section 6 presents the results and performs

comparative analysis between the VSM and LSA. Finally, the most important conclusions are outlined in Section 7.

II. Related Work

Commercially available conference management systems usually rely on explicit methods of describing papers and reviewers' competencies, most commonly selection of keywords/topics from a predefined list or a taxonomy [2]. However, in the recent years some of them started to implement more complex IR approaches, performing text analysis of the submitted papers and the previous reviewers' publications.

Pesenhofer et al. [3] suggest that paper-reviewer similarities are calculated as Euclidian distance between the titles of the submitted papers and the titles of all reviewers' publications. The authors evaluated their approach with data from ECDL 2005. They noted that for 10 out of 87 PC members, no publications have been found and they got their papers to review at random.

Ferilli et al. [4] use Latent Semantic Indexing (LSI, LSA) to identify reviewers to evaluate submitted papers. The document collection consisted of the titles and the abstracts of the submitted papers and the titles of reviewers' publications obtained from DBLP. Results were evaluated by the organizers of the IEA/AIE 2005 conference. In their opinion the average accuracy was 79%. According to the reviewers, the accuracy was 65% [4].

Charlin and Zemel [5],[6] propose a standalone paper assignment recommender system called "The Toronto Paper Matching System (TPMS)". It builds reviewers' profiles based on their previous publications obtained from Google Scholar or uploaded by the reviewers themselves. By using Latent Dirichlet Allocation (LDA)[1], TPMS extracts reviewers' research topics from their publications.

Dumais and Nielsen [8] used latent semantic indexing to automate the assignment of papers to reviewers in Hypertext'91 conference. Their results show a mean number of relevant articles in the top-10 of 5.9, and average precision value of 0.51. They conclude that the simple LSI method is not as good as the best human experts, but it could perform in the same general range and achieves the same performance as a human, who is not a narrow expert in the field, but has broader view and good knowledge in it [8].

Moldovan et al. [9] compare the performance of latent semantic analysis to the vector space model (VSM) applied to US patent documents from 1790 to 2005. Their results show that LSA almost always matches the VSM and sometimes slightly outperform it with an average improvement of 5%, and in a single case it performed worse with an average damage of 3% [9]. It should be noted that they were not using any term weighting model in the term-document matrix.

Many researchers (Nguyen et al. [10], Liu et al. [11], Conry et al. [12]) are proposing more complex composite methods to identify proper reviewers to evaluate papers, that also applies content analysis and IR approaches (especially LDA) on multiple data sources, not just publications' abstracts. Liu et al. [11] suggest that paper-reviewer similarities are calculated based on three aspects of the reviewer, which are lately integrated by a Random Walk with Restart (RWR) model. Authors compare their approach to other IR techniques like "text similarity" (i.e. VSM) and "topic similarity" (derived by LDA) and more or less surprisingly, their results show that text similarity actually outperforms topic similarity. So, pure VSM with proper term-weighing model could sometimes perform better than topics extraction by LDA followed by a cosine similarity of the topic vectors.

III. Using Vector Space Model (VSM) to Identify Competent Reviewers

According to the vector space model, the meaning of a document is obtained from its words. Thus, the document could be represented by an array (vector) of words. Not just its words, but all unique words from the entire document collection. This provides equal length of all document vectors and allows easy calculation of similarity between two documents by using cosine similarity. However, in case of large document collections, the vectors' length could get enormous (with most elements set to 0) that makes calculation of similarities ineffective. Fortunately, this could be overcome by using inverted index instead of forming document vectors with tens of thousands dimensions.

Document vectors do not actually contain the words (terms) themselves, but their weight instead. There are many ways of calculating term weight (called term-weighing models), but they are all based on two main components: term frequency (tf) - the number of occurrences of a term ti in the document dj; and document frequency (df) - the number of documents that contain ti. The presumption is that the more times a term occur in a document, the more important it is for that document. But, the more documents contain a term, the less informative it is. As df is an inverse measure of informativeness, we use not df, but idf – inverse document frequency. The most basic term-weighting model is the simple multiplication of tf * idf. However, there are more complex and accurate models that rely on compositions of different tf normalization functions – Singhal [13], Rousseau and Vazirgiannis [14], Robertson's BM 25 [15],[16] and others. Comparison of these models in the context of reviewer assignment problem could be found in [17]. Once terms weights are calculated, the similarity between two documents (or between the query and a document) could be easily calculated as the cosine of the angle between the two vectors.

A comprehensive experimental analysis aiming to check if the VSM could be reliably used for automatic identification of proper (competent) reviewers to evaluate papers is performed in my previous work [17]. According to the results, the short answer is "yes, it could be". It produces 5-10% less accuracy in comparison with the explicit selection of keywords from a taxonomy, but still high enough accuracy that allows the VSM to be used as a stand-alone method. Results also show that:

- The Robertson's BM 25 weighing model [15] achieves highest accuracy.
- Word stemming further increases identification accuracy.

- Using IDF only on query terms, rather than on both – the query and the documents terms, provides better results.
- Complex term-weighting models that consist of composition of different TF normalization functions provide better results than the plain lnc.ltc scheme.

Experiments were performed on real datasets, taken from the CompSysTech series of conferences for a 5 years period – from 2014 to 2018. Experiments in this study are using absolutely the same input datasets and evaluations methods, so a fair and objective comparison could be done between the vector space model and the latent semantic analysis in the context of reviewer assignment problem.

IV. Using Latent Semantic Analysis (LSA) to Identify Competent Reviewers

The latent semantic analysis (LSA) is a dimension reduction (or rank lowering) technique applied over the bag-of-words (BoW) model, that analyzes relationships between documents, but also relationships between the words they contain. The latter is very important and a major difference from the VSM. The assumption is that words which have similar meanings often occur in the same documents. Thus LSA is able to "group" semantically-related words into broader topics. The method is called "latent semantic analysis" because it discovers a number of latent (hidden) topics that could describe (separately or in combination) each document within the collection. These topics are not the exact words but they have more generalized meaning. Each word/term in the collection's dictionary is related or has a specific contribution to some topic(s). Similarly, each topic has a specific contribution to some documents. For example, the words: space, booster, shuttle, rocket, probe form the "space-related" topic. A document could be related to space if it contains any of these words. In contrast, if we apply the VSM over BoW, then each term is treated separately. For example shuttle and rocket are entirely dissimilar. However for the LSA, these terms are related. In this sense, LSA can cope with synonyms and partly with polysemy that is a great advantage in comparison to VSM. So in theory, word stemming is not necessary in preprocessing. But it will be tested during the experiments.

The input of the LSA is the term-document matrix (the leftmost part of Fig. 1) – a matrix where rows represent terms and columns represent documents. Generally, it states how many times each document contain each term, but values could be also the tf-idf weights of the terms in respect to each document.

Let's call the term-document matrix A. The row ai contains the weights of the i-th term in respect to all documents. Similarly, the row ap represents the p-th term. The dot product $a_i^T a_p$ indicates how related the i-th and the p-th terms are. Applying cosine normalization of the dot product, we get the cosine similarity between these two terms.

The matrix product AA^T will contain similarity factors between all terms in the entire document collection. Similarly, calculation of $A^T A$ provides similarities between all documents for the entire dictionary.

There is a matrix factorization technique in the linear algebra, called Singular Value Decomposition (SVD). According to it a matrix could be decomposed in three matrices such that:

$$A = U\Sigma V^T \quad (1)$$

where U and V are orthogonal matrices, and Σ is a diagonal matrix. The values in Σ are called singular values and they show the significance of each latent topic. Values are ordered in the main diagonal in descending order, placing the most significant topic on top. The values of U are called left singular values and they indicate the contribution of each term to each discovered topic. The values of V^T are called right singular values and show the contribution of each topic to each document. The idea is illustrated with 4 terms, 3 documents (A4x3) and 2 topics on Fig. 1.

It should be noticed that in general, if A has a dimension of mxn, then the dimension of U is mxm, the dimension of Σ is mxn and dimension of V^T is nxn. However, as LSA is a dimension reduction technique, only the highest k singular values (the k most significant topics) and their corresponding singular vectors form U and V are taken into account, performing a truncated SVD. Then, as in the example above, the dimension of U is mxk, the dimension of Σ is kxk and dimension of V^T is kxn.

It could be proven that the columns of U are actually the eigenvectors of the matrix product AA^T, the columns of V (or rows of V^T) are the eigenvectors of A^TA, and the singular values of Σ are the square roots of the eigenvalues of AA^T or A^TA.

Thus, calculating SVD requires calculation of the eigenvalues and the eigenvectors of the matrix products AA^T and A^TA.

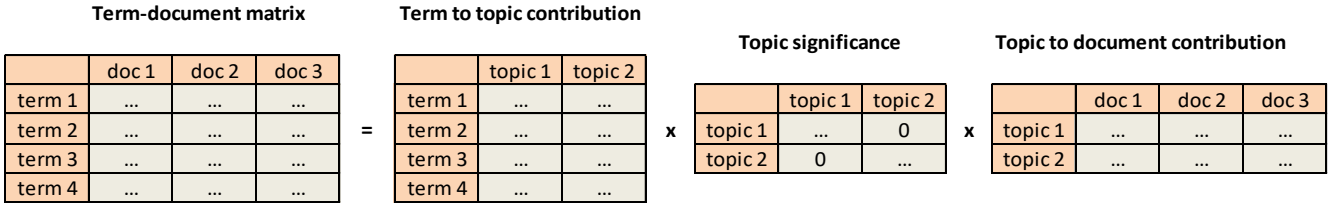

Fig. 1. SVD Decomposition of the Term-Document Matrix into Three Matrices, providing the Significance of each Latent Topic and the Contribution of Terms to Topics and Topics to Documents.

From the linear algebra, it is also known that

$$Av = \lambda v \quad (2)$$

where A is a matrix, v is an eigenvector and λ is an eigenvalue of the matrix. Equation (2) is called eigenvalue equation of A.

Equation (2) could be rewritten in the form of

$$(A - \lambda I)v = 0 \quad (3)$$

where I is the identity matrix.

Eq. (3) could have a non-zero solution (eigenvector) v, if the determinant of the matrix $(A - \lambda I)$ is zero. Thus:

$$|A - \lambda I| = 0 \quad (4)$$

So, the eigenvalues are calculated by the characteristic equation (4). The determinant of it is a polynomial function of λ with degree n, where n is the order of A. Thus the characteristic equation (4) has up to n solutions for λ which are the eigenvalues of A.

After calculating the eigenvalues of A, the eigenvector that corresponds to each eigenvalue could be determined by solving the linear equation system (3). Then the eigenvectors of AA^T form the columns of U and the eigenvectors of A^TA form the columns of V^T.

Finally, calculating the similarity between two documents in the lower dimensional k space, means calculating the cosine similarity between their corresponding columns of V^T.

If the query is missing in V^T, the original query vector should be transformed to the lower dimensional k space first (eq. 5), then the transformed query qk could be compared (by cosine similarity) with any document from V^T.

$$q_k = \Sigma_k^{-1} U_k^T q \quad (5)$$

V. Experimental Setup

Testing whether the Latent Semantic Analysis (LSA) could be successfully and reliably used to identify experts to review submitted papers is done by using real datasets taken from already conducted conferences for a period of 5 years – CompSysTech [18] from 2014 to 2018. The same datasets are used in the previous study aiming to test if the Vector Space Model (VSM) and the explicit selection of keywords from a taxonomy could be used for the same purpose. That allows completely fair and objective comparison between all these three methods of reviewer identification. All datasets contain reference values (the ground truth) for the level of competency of each reviewer in each paper he/she evaluates. These values are explicitly stated by the reviewers themselves during the review submission. So they could be considered as 100% accurate and used as a reference (benchmarks).

The document collection consists of the titles and the abstracts of all submitted papers and the titles and the abstracts of all reviewers' previous publications. The former are taken directly from the CompSysTech database, while the latter are fetched from the joint API of DBLP and Semantic Scholar. It allows getting data (full bibliography, including abstracts) from Semantic Scholar while searching by name in the DBLP's database.

Before applying the latent semantic analysis, the content of all documents is preprocessed as follows:

1) All punctuation marks (commas, dots, dashes, exclamation, quotation and question marks and etc.) are removed since they only cause troubles. If they stick to the words, that makes term recognition harder (for example "red" and "red," will be recognized as two different terms, because of the comma). If they are separated with spaces, however, they could be recognized as terms, making document vectors longer and decreasing the relative weight of meaningful terms.

2) All the text is converted to lowercase. This makes the analysis case-insensitive.

3) The text is tokenized. This is the process of splitting the text into an array (vector) of terms.

4) All semantically-insignificant terms (so called "stop words") are removed. These are prepositions, conjunctions, pronouns and etc. They are important from a syntactic point of view, but they do not represent any semantic, meaning and subject domain of the documents. Furthermore, as they are frequently appearing anywhere in the text, they will have disproportionally high tf value in comparison to the semantically-meaningful words, i.e. the semantically-insignificant stop words will highly lower the relative weight of the semantically-significant ones, which is undesirable. That's why stop words should be removed. Stop words are usually pre-defined as a list or array, and of course, they are language-dependent.

5) Finally, the Porter's word stemming algorithm [19] is applied on all remaining tokens. This is an optional step and could be skipped. Word stemming is the process of separation of word endings from the morphological root. The idea is to keep and process just the roots and skip word endings. In this way, different forms of a single word (for example: beautiful, beauty, beautifully) could be recognized as one.

The very same preprocessing is applied in the previous study [17] of the possibility of using VSM to identify reviewers. So, again, both methods are tested using the same preprocessing activities and with the same input data.

The ultimate goal of the LSA is to calculate a similarity factor between every submitted paper and every registered reviewer (PC member). It shows how competent the reviewer is to evaluate the specified paper. However reviewers have more than one publication in their profiles. Thus, a similarity factor is calculated between every submitted paper and every reviewer's publication. Then the overall similarity between a paper and a reviewer is summarized as an average of the 10% highest similarity factors between the paper and the reviewer's publications. However, the 10% number of reviewer's publications taken into account (in the overall similarity) could not be less than 3.

When performing the experiments, there are some very important settings whose value could highly impact the LSA's accuracy. These are:

- The number of latent (hidden) topics.
- The way the term-document matrix is formed. Whether it contains raw number of term occurrences or tf-idf normalized values.
- The term-weighing models in case of tf-idf normalized matrix.
- Whether word stemming is applied or not.

The number of latent topics is probably the most important setting but there is no theoretically-motivated correct value. It should be experimentally determined. Using too many topics may cause the LSA to behave like the vector space model, treating terms separately. However, choosing too few topics will cause the LSA to group unrelated terms together, losing accuracy.

For a raw term-document matrix of collection of about 5000 documents, the experiments started with 100 topics as previous research by other scientists [8], [9], [20] suggest it is a good starting point. If the number of documents and unique terms gets lower (or higher), then the number of latent topics should be decreased (increased) as well. That assumption is fully supported by the experiments in this work as well.

In general, the LSA uses a term-document matrix containing raw values, i.e. just the number of occurrences of each term in each document. However, the experiments in this article show significant increase in accuracy when the term-document matrix is composed of tf-idf term weights, rather than just the raw number of occurrences.

Experiments are performed on custom software developed in php and Matlab. The php part is responsible for extracting reviewers' publications from the Internet, building the document collection and exporting it within a proper structure in text files. The LSA is implemented in Matlab since it has a built-in function to perform the SVD decomposition.

VI. Experimental Results

To determine if the paper-reviewer similarity factors, obtained by LSA, VSM or other method are correctly calculated, they have to be compared to some reference evaluation of expertise that we trust it is correct. Since real datasets are used for experimental evaluation, fortunately, there is such a reference. During review submission, reviewers are required to explicitly indicate their level of expertise (High, Medium or Low) in respect to each paper they evaluate. As the reviewers themselves explicitly provide these levels, it could be assumed they are completely accurate and they could be used as a reference. However, the two data values (paper-reviewer similarity factors and levels of expertise) are not directly comparable. Similarity factors are decimals within the range [0.00, 1.00], while the explicitly stated levels of expertise are just "labels" – low, medium and high. To overcome this problem, a special-purpose software has been developed that converts similarity factors to levels of expertise, and then performs a correlation analysis between the automatically determined levels of expertise and the ones explicitly stated by the reviewers during the review submission. The conversion is done based on the assumption that if a reviewer ri has declared higher level of expertise than another reviewer rj (for the same paper), then ri should has higher similarity factor with the paper than rj. Detailed description of the software could be found in [21]. It is used to evaluate all the three methods – the latent semantic analysis, the vector space model and the selection of keywords from the conference taxonomy. So, again, they are all placed on equal terms (share the same input data and evaluation method) and thus could be objectively compared.

Experiments started with the data from the CompSysTech 2018 conference. Initial experiments aimed to test the influence of the previously mentioned factors – the number of latent topics, the term-weighting models and word stemming. Accuracy of the computed similarity factors is evaluated by the percentage of the correctly calculated similarities and their correlation with the levels of expertise, explicitly stated by the reviewers themselves during review submission. A similarity factor is considered to be correctly calculated, if it complies with the rules stated in [21].

The CompSysTech 2018 dataset consists of 75 submitted papers and 73 registered reviewers. After adding the abstracts of all reviewers' previous publications, the entire document collection became 4648 documents, having 21 682 unique words.

A. Experiment 1: Testing if the Number of Hidden Topics Influence the Accuracy of the Calculated Similarity Factors

The term-document matrix contains raw term frequencies, i.e. the number of occurrences of each term in each document. No stemming is applied.

As expected, results show that the number of latent topics indeed influences the accuracy of the calculated paper-reviewer similarities. Moreover, the experiment also confirm that if the document collection consists of about 5K documents and the term-document matrix contains raw tf values, then 100 is the optimal number of latent topics to start with.

B. Experiment 2: Testing if the Term-Weighting Models Influence the Accuracy of the Calculated Paper-Reviewer Similarities

In the vector space model (VSM), composite and more complex term-weighting schemas usually achieve higher accuracy than using the raw number of term occurrences. It is curious to test if this fact is valid for the LSA as well. It should be. So in this experiment, the term-document matrix does not contain the raw term frequencies (as in experiment 1), but the term weights are calculated by the basic TF-IDF model (6). Two series of experiments were performed, first with IDF applied on both the document terms and the query terms, and then with IDF applied only on query terms. TF stands for term frequency, while IDF for inverse document frequency. For more information, please refer to [17].

$$w_{i,j} = \left(1 + \log(tf_{i,j})\right) * \log(\frac{d}{df_i}) \quad (6)$$

Comparing Table I and Table II, it is clearly noticeable that the accuracy of paper-reviewer similarities gets significantly

higher when the term-document matrix is composed of TF-IDF term weights, rather than the raw number of term appearances.

TABLE I. PERCENTAGE OF CORRECTLY CALCULATED PAPER-REVIEWER SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH THE EXPLICIT REVIEWERS' OPINION FOR COMPSYSTECH 2018. TERM-DOCUMENT MATRIX CONTAINS RAW NUMBER OF TERM OCCURRENCES. NO STEMMING IS APPLIED

# latent topics	50	75	100	125	150
% correctly calculated	71.36 %	72.27 %	**75 %**	74.55 %	74.09 %
Pearson correlation	0.6254	0.6423	**0.6669**	0.6610	0.6588

TABLE II. PERCENTAGE OF CORRECTLY CALCULATED PAPER-REVIEWER SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH THE EXPLICIT REVIEWERS' OPINION FOR COMPSYSTECH 2018 AT DIFFERENT NUMBER OF LATENT TOPICS AND TERM WEIGHTING MODELS. NO STEMMING

# latent topics	10	20	30	40	50	75
Weighting model: basic TF-IDF (eq. 6), IDF applied on both document and query terms (ltc.ltc)						
% correctly calculated	79.09	79.55	**82.73**	81.36	80.91	76.82
Pearson correlation	0.7037	0.7100	**0.7610**	0.7417	0.7332	0.6819
Weighting model: basic TF-IDF (eq. 6), IDF applied only on query terms (lnc.ltc)						
% correctly calculated	75	76.82	**77.27**	74.55	74.55	75
Pearson correlation	0.6743	0.6841	**0.6867**	0.6521	0.6521	0.6513

Many researchers have proven that in VSM it is better to skip IDF for document terms and apply it just on query terms. This makes a lot of sense since a frequently appearing term in a document says it (the term) is important for the document semantics. However if IDF is applied on it, that may significantly reduce its weight, making it semantically insignificant (which is not the case). Experimental results in Table II; however, show this sense is not applicable to LSA and skipping IDF for document terms does not improve, but actually worsens accuracy.

Another interesting observation in Table II is that higher accuracy is achieved in lower number of hidden topics. This is also important since lower number of latent topics means lower dimension of the SVD transformation matrices, thus lower computational complexity and lower execution time.

C. Experiment 3: Checking if Word Stemming could Increase Accuracy

Word stemming increases accuracy in the vector space model since it recognizes different forms of a single word (for example: beautiful, beauty, beautifully) as one. However, in case of latent semantic analysis it should have minimal or no effect, because the basic idea of LSA is to group words with similar meaning together, making word stemming unnecessary. The aim of this experiment is to check this assumption.

Word stemming in this experiment is done by Poster's stemming algorithm [19] before constructing the term-document matrix. For more reliability and determination, it is tested with both the raw number of term occurrences and the TF-IDF weighting model (Table III).

TABLE III. PERCENTAGE OF CORRECTLY CALCULATED SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH REVIEWERS' OPINION FOR COMPSYSTECH 2018 WITH PORTER STEMMER APPLIED BEFORE LSA

Porter Stemmer, weighting model: raw number of term occurrences					
# latent topics	50	100	130	150	170
% correctly calculated	75	75	75.45	75.45	74.55
Pearson correlation	0.6585	0.6625	0.6614	0.6689	0.6564
Porter Stemmer, weighting model: TF-IDF (ltc.ltc)					
# latent topics	10	20	30	40	50	75
% correctly calculated	79.09	80.45	**81.82**	80	80	78.2
Pearson correlation	0.7192	0.7235	**0.7413**	0.7151	0.7213	0.70

A brief look at experiment 1 shows that without word stemming, 75% of similarity factors are correctly calculated and the correlation with reviewers' opinion is 0.6669. With a word stemming, correctly calculated similarities are 75.45% and the correlation is 0.6689. So, as expected, word stemming has an insignificant (negligible) impact on the accuracy. The combination of word stemming with TF-IDF weighted term-document matrix even lowers accuracy a little bit.

To summarize experiments 1 to 3, it can be concluded that highest accuracy (percentage of correctly calculated similarity factors and correlation with reviewers' opinion) is achieved when no word stemming is applied, and the term-document matrix is composed of TF-IDF weights, rather than raw number of term occurrences. The number of hidden (latent) topics greatly affects the accuracy as well, but it is also depends on the number of documents and unique words (terms) within the document collection, so an exact number could not be defined in advanced.

Another assumption is experimentally proven as well - that lowering the number of latent topics, increases the value of the calculated similarity factors. This is expected, but higher values of all computed similarities do not mean they are accurately calculated and real-life paper-reviewer similarities are high as well. So, the number of hidden topics should not be lowered too much or it may highly distort the results. Experiments show that in case of TF-IDF weighted term-document matrix, going down to 5 or less topics, produces very high values (>0.9) for all paper-reviewer similarities, which of course cannot be true.

D. Experiment 4: Testing LSA with other CompSysTech Datasets

To verify that results for CompSysTech 2018 are not obtained by a lucky chance, the latent semantic analysis is applied (without word stemming) on all CompSysTech issues for a 5 year period of time – from 2014 to 2018.

It should be noted here that when downloading the abstracts of reviewers' previous publications, only manuscripts published before the specific conference year are taken into account. For example, if the conference is in 2015, then

reviewers' publications up to 2014 (including) are considered for processing. For that reason the document collection of CompSysTech 2014 will be smaller than the one of 2018, regardless of the number of actual reviewers.

The percentage of correctly calculated paper-reviewer similarities and the level of their correlation with the reviewers' opinions for the other CompSysTech issues (2017 to 2014) are presented in Tables IV to VII. As expected, the highest accuracy (marked in green) in all cases is obtained with TF-IDF weighted term-document matrix. However, it could be seen that going back in time, it is achieved at lower number of latent topics – 30 for CompSysTech 2018, and just 15 for CompSysTech 2014, 2015 and 2016. That is expected and pretty logical – as we go back in time, the document collection gets smaller (from 4648 to 2550 documents) due to the lower number of reviewers' publications. The smaller the document collection, the lower is the number of unique words, leading to lower optimal number of hidden topics.

TABLE IV. PERCENTAGE OF CORRECTLY CALCULATED PAPER-REVIEWER SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH THE EXPLICIT REVIEWERS' OPINION FOR COMPSYSTECH 2017 AT DIFFERENT NUMBER OF LATENT TOPICS AND TERM-DOCUMENT MATRIX'S WEIGHTING MODELS

Papers: 107, Reviewers: 76 Documents: 4128, Unique words: 19698					
Weighting model: raw term frequencies					
# latent topics	80	100	120	140	
% correctly calculated	76.19	76.83	77.46	75.56	
Pearson correlation	0.7482	0.7579	0.7647	0.7460	
Weighting model: basic TF-IDF (eq. 6), ltc.ltc					
# latent topics	10	20	25	30	40
% correctly calculated	80	81.59	**81.9**	80.95	80.63
Pearson correlation	0.7867	0.8019	**0.8078**	0.7980	0.7916

TABLE V. PERCENTAGE OF CORRECTLY CALCULATED PAPER-REVIEWER SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH THE EXPLICIT REVIEWERS' OPINION FOR COMPSYSTECH 2016 AT DIFFERENT NUMBER OF LATENT TOPICS AND TERM-DOCUMENT MATRIX'S WEIGHTING MODELS

Papers: 117, Reviewers: 73 Documents: 3926, Unique words: 19787				
Weighting model: raw term frequencies				
# latent topics	60	80	100	120
% correctly calculated	73.93 %	74.79 %	74.5 %	73.93 %
Pearson correlation	0.6956	0.7056	0.6974	0.6876
Weighting model: basic TF-IDF (eq. 6), ltc.ltc				
# latent topics	15	20	25	30
% correctly calculated	**80.8 %**	80.52 %	79.66 %	78.22 %
Pearson correlation	**0.7500**	0.7451	0.7414	0.7250

TABLE VI. PERCENTAGE OF CORRECTLY CALCULATED PAPER-REVIEWER SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH THE EXPLICIT REVIEWERS' OPINION FOR COMPSYSTECH 2015 AT DIFFERENT NUMBER OF LATENT TOPICS AND TERM-DOCUMENT MATRIX'S WEIGHTING MODELS

Papers: 103, Reviewers: 74 Documents: 3090, Unique words: 17165				
Weighting model: raw term frequencies				
# latent topics	40	50	60	80
% correctly calculated	76.95 %	76.95 %	76.27 %	75.25 %
Pearson correlation	0.7280	0.7383	0.7388	0.7213
Weighting model: basic TF-IDF (eq. 6), ltc.ltc				
# latent topics	8	10	15	20
% correctly calculated	81.69 %	82.71 %	**82.71 %**	81.36 %
Pearson correlation	0.7720	0.7842	**0.7874**	0.7717

TABLE VII. PERCENTAGE OF CORRECTLY CALCULATED PAPER-REVIEWER SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH THE EXPLICIT REVIEWERS' OPINION FOR COMPSYSTECH 2014 AT DIFFERENT NUMBER OF LATENT TOPICS AND TERM-DOCUMENT MATRIX'S WEIGHTING MODELS

Papers: 107, Reviewers: 65 Documents: 2550, Unique words: 14810				
Weighting model: raw term frequencies				
# latent topics	30	40	60	80
% correctly calculated	74.1 %	74.75 %	73.77 %	71.8 %
Pearson correlation	0.6791	0.6835	0.6756	0.6574
Weighting model: basic TF-IDF (eq. 6), ltc.ltc				
# latent topics	8	10	15	20
% correctly calculated	78.69 %	78.69 %	**79.67 %**	78.36 %
Pearson correlation	0.7114	0.7074	**0.7340**	0.7137

Finally, it is interesting to see a direct performance comparison between the latent semantic analysis (LSA), the vector space model (VSM) and the explicit document description by a taxonomy of keywords in computing paper-reviewer similarities for all issues of CompSysTech. Such a comparison is presented in Table VIII. It includes only the highest accuracies, obtained by the VSM and LSA for every CompSysTech issue. Data for the VSM and the method of describing papers/reviewers by taxonomy of keywords are taken from my previous publication [17]. All methods are tested by using the same input data (CompSysTech 2014-2018 datasets) and by the same similarity factors' evaluation tool [21]. So, comparison is fair and objective.

There are dozens of experiments, testing many popular TF-IDF weighting models with the VSM in [17]. However, Table VIII shows only the best performing one – the algebraic version of Robertson's BM 25.

Expectedly, LSA outperforms VSM, even for the best performing term-weighting model for VSM. However, it is a bit surprising that, in some cases, LSA slightly outperforms the explicit document description by taxonomy of keywords as well.

TABLE VIII. COMPARISON OF LSA, VSM AND THE EXPLICIT DOCUMENT DESCRIPTION BY A TAXONOMY OF KEYWORDS FOR ALL COMPSYSTECH ISSUES FROM 2014 TO 2018

	CST 2018		**CST 2017**		**CST 2016**		**CST 2015***		**CST 2014***	
Total assignments	220		315		349		295		305	
	Correctly calculated, %	Correlation	Correctly calculated, %	Correlation	Correctly calculated, %	Correlation	Correctly calculated, %	Correlation	Correctly calculated, %	Correlation
Taxonomy of keywords [2]										
	81.82	0.75	81.90	0.80	80.23	0.74	**85.08**	**0.81**	**80.66**	**0.78**
Vector Space Model (VSM), Robertson's BM25 TF_{kep} / TF_{kep} x IDF										
No stemming	73.64	0.67	76.51	0.74	72.78	0.65	75.59	0.70	72.13	0.67
Porter stemmer	74.09	0.66	79.37	0.77	73.64	0.68	76.95	0.72	74.43	0.68
Latent Semantic Analysis (LSA), TF-IDF weighted term-document matrix, eq. (6)										
No stemming	**82.73**	**0.76**	**81.90**	**0.81**	**80.80**	**0.75**	82.71	0.79	79.67	0.73

* Three PC members of CompSysTech 2015 and 2014 were not identifiable in DBLP.

It should be noted here, that 3 PC members of CompSysTech 2014 and 2015 were not found in DBLP, so the abstracts of their previous publications were excluded from the document collection, meaning they get zero similarities with all papers. Actually, missing data for some reviewers is the highest threat to LSA and VSM since they calculate similarities based on content analysis. If there is no content, there is no similarity, and those reviewers could have their papers assigned at random.

Results of the LSA to VSM comparison comply with most of the previous similar research. Although the LSA achieves an increase of 30% in the average accuracy for the MED collection, it shows much lower improvement for CISI and NPL datasets, while performing even worse for TIME and CACM collections [22]. In real-life applications, improvement is also moderate. Moldovan et al. [9] applied both LSA and VSM to analyze US patent documents and their results show that LSA slightly outperform VSM with an average improvement of up to 5%. That is fully comparable to the results obtained in this study, in case the term-document matrix is composed of raw term frequencies. However, if the term-document matrix is composed of tf-idf weights, accuracy could be increased with up to 10% in respect to the VSM.

VII. CONCLUSION

After performing large number of experiments with all the five CompSysTech datasets, it can be concluded that:

1) The latent semantic analysis (LSA) could be accurately and reliably used to identify competent reviewers to evaluate papers.

2) The latent semantic analysis outperforms the vectors space model in almost all cases, even when VSM implies the Robertson's BM 25 as a term-weighting model.

3) When the term-document matrix of LSA is composed of raw number of term occurrences, the LSA slightly outperforms VSM by 2-3 ppts (percentage points).

4) Composing the term-document matrix of TF-IDF weights, rather than raw number of term occurrences, additionally boosts accuracy by further 5 ppts, and allows the LSA even to slightly outperform the method of explicit document description by a taxonomy of keywords.

5) In contrast to the vector space model, the LSA achieves higher accuracy when IDF is applied to both document and query terms.

6) Word stemming has a little effect on accuracy of similarities computed by LSA.

7) The optimal number of latent (hidden) topics depends on the number of unique words (terms) within the document collection. Higher number of terms results in higher optimal number of latent topics, and the opposite.

8) Lowering the number of latent topics increases the values of all calculated paper-reviewer similarities, but not their accuracy.

9) The highest threat in using LSA to assign reviewers to papers is to have a PC member who cannot be found in DBLP and Semantic Scholar. In this case, no publications could be extracted for him/her and he/she will get zero similarities with all papers. The latter means that papers will be assigned to him/her at random.

Both the latent semantic analysis and the vector space model could be reliably used to identify competent reviewers to evaluate papers. LSA achieves higher accuracy, but it is harder to be implemented and has higher time complexity. Furthermore, in contrast to the VSM, the LSA could not be computed by using an inverted index, making it much slower than VSM. Additionally, the accuracy of LSA depends on the number of latent topics, but the optimal number could not be

set in advanced. So, although LSA achieves higher accuracy, the VSM may be a better choice for commercially available conference management systems due to its simplicity and better time complexity, allowing real-time computation even for large scale conferences.

Other IR approaches (most probably composition of several methods and/or data sources) will be tested in future to check if they could also be used to identify competent reviewers to evaluate submitted papers. So far, both the VSM and the LSA, together with the method of explicit description of papers and reviewers by choosing keywords from a predefined taxonomy, turned to be quite reliable option for this task.

REFERENCES

[1] Y. Kalmukov, "An algorithm for automatic assignment of reviewers to papers", Scientometrics, 2020, No 124 (3), pp. 1811–1850, https://doi.org/10.1007/s11192-020-03519-0.

[2] Y. Kalmukov, "Describing Papers and Reviewers' Competences by Taxonomy of Keywords", Computer Science and Information Systems, 2012, No 9(2), pp. 763-789, https://doi.org/10.2298/CSIS110906012K.

[3] A. Pesenhofer, R. Mayer, A. Rauber, "Improving Scientific Conferences by enhancing Conference Management System with information mining capabilities", Proceedings IEEE International Conference on Digital Information Management (ICDIM 2006), ISBN: 1-4244-0682-x; S. 359 - 366.

[4] S. Ferilli, N. Di Mauro, T.M.A. Basile, F. Esposito, M. Biba, "Automatic Topics Identification for Reviewer Assignment", 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2006. Springer LNCS, 2006, pp. 721-730.

[5] L. Charlin and R. Zemel, "The Toronto paper matching system: an automated paper-reviewer assignment system." (2013).

[6] L. Charlin, R. Zemel and C. Boutilier, "A framework for optimizing paper matching", In Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (Corvallis, OR, 2011). AUAI Press, 86–95.

[7] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation", Journal of Machine Learning Research 2003, 3:993-1022.

[8] Susan T. Dumais, and Jakob Nielsen, "Automating the assignment of submitted manuscripts to reviewers." In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 233-244. 1992.

[9] Andreea Moldovan, Radu Ioan Bot and Gert Wanka, "Latent semantic indexing for patent documents." (2005).

[10] Jennifer Nguyen, Germán Sánchez-Hernández, Núria Agell, Xari Rovira, and Cecilio Angulo, "A decision support tool using Order Weighted Averaging for conference review assignment", Pattern Recognition Letters 105 (2018): 114-120.

[11] Xiang Liu, Torsten Suel and Nasir Memon. "A robust model for paper reviewer assignment", In Proceedings of the 8th ACM Conference on Recommender systems, pp. 25-32. 2014.

[12] Don Conry, Yehuda Koren, and Naren Ramakrishnan. "Recommender systems for the conference paper assignment problem", In Proceedings of the third ACM conference on Recommender systems, pp. 357-360. 2009.

[13] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization", In Proceedings of SIGIR'96, pages 21–29, 1996.

[14] F. Rousseau and M. Vazirgiannis, "Composition of TF normalizations: new insights on scoring functions for ad hoc IR", In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 917-920. 2013.

[15] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF", Journal of documentation, vol. 60 no. 5, pp 503–520, 2004.

[16] S. E. Robertson, S. Walker, K. Spärck Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3", In Proceedings of TREC-3, pages 109–126, 1994.

[17] Y. Kalmukov, "Automatic Assignment of Reviewers to Papers Based on Vector Space Text Analysis Model", Proceedings of the 21st International Conference on Computer Systems and Technologies, CompSysTech 2020, Association for Computing Machinery, New York, NY, USA, 2020, pp. 229–235, DOI: https://doi.org/10.1145/3407982.3408026.

[18] CompSysTech – International conference on computer systems and technologies, http://www.compsystech.org.

[19] Matrin F. Porter, "An algorithm for suffix stripping", In J. S. Karen and P. Willet, editors, Readings in information retrieval, pages 313-316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[20] D. Grossman, O. Frieder, "Information Retrieval: Algorithms and Heuristics", 2nd Ed. Springer, The Netherlands, 2004, ISBN: 1-4020-3004-5.

[21] Y. Kalmukov, "A Software Tool for Accuracy Evaluation of Calculated Similarity Factors between Papers and Reviewers", 2020 7th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE), 2020, pp. 1-5, IEEE, https://doi.org/10.1109/EEAE49144.2020.9279032.

[22] Dandan Li and Chung - Ping Kwong, "Understanding latent semantic indexing: A topological structure analysis using Q - analysis", Journal of the american society for information science and technology 61, no. 3 (2010): 592-608.