PRIMA - Towards an Automatic Review/Paper Matching Score Calculation

Christian Caldera, René Berndt, Eva Eggeling
Fraunhofer Austria Research GmbH
Email: {christian.caldera, rene.berndt, eva.eggeling}
@fraunhofer.at

Martin Schröttner
Institute of Computer Graphics and Knowledge Visualization
University of Technology, Graz, Austria
Email: martin.schroettner@cgv.tugraz.at

Dieter W. Fellner
Institute of ComputerGraphics and KnowledgeVisualization (CGV), TU Graz, Austria
GRIS, TU Darmstadt & Fraunhofer IGD, Darmstadt, Germany
Email: d.fellner@igd.fraunhofer.de

***Abstract*—Programme chairs of scientific conferences face a tremendous time pressure. One of the most time-consuming steps during the conference workflow is assigning members of the international programme committee (IPC) to the received submissions. Finding the best-suited persons for reviewing strongly depends on how the paper matches the expertise of each IPC member. While various approaches like "bidding" or "topic matching" exist in order to make the knowledge of these expertises explicit, these approaches allocate a considerable amount of resources on the IPC member side. This paper introduces the *Paper Rating and IPC Matching* Tool (PRIMA), which reduces the workload for both - IPC members and chairs - to support and improve the assignment process.**

***Keywords*-Conferences, International Program Committee, Submissions, Paper, Assignment, Matching, TF-IDF, Information Retrieval.**

I. INTRODUCTION

Conferences and journals play an important role in the scientific world. Both are important channels for exchange of information between researchers. The publication list of a researcher defines his/her standing within the scientific community. In order to ensure quality standards for these publications, submitted work undergo the so called peer review process. This process is used to maintain standards, improve performance and provide credibility [1]. Today almost every conference or journal uses a electronic conference management system in order to organize this process.

In-a-nutshell the peer review process for a conference works as follows: First, authors upload their paper to the electronic submission system. After the deadline the submitted papers are distributed to the reviewers. The conference reviewers are usually members of the International Programme Committee (IPC) and - depending on the size of the conference - a pool of external experts. Each reviewer receives a certain amount of submitted papers depending on his/her expertise. Assigning the submitted papers to the IPC members is a crucial task in the peer review process because these reviews decide if the paper is accepted or not. In case of acceptance the author is allowed to upload a camera-ready-copy version of the paper, which is then published in the proceeding of the conference.

This is the essence of the peer review process. Within the peer review process there exist variations, which mostly differ in what information is revealed to whom. The most commonly used are the single and double blinded peer reviewing process:

- In the single blinded peer review the identity of the reviewer is unknown to the user. But, the reviewer knows the identity of the author. In this setting, the reviewer can give a critical review without the fear that the person itself will be targeted by the author.
- In the double blinded peer review, the identity of the reviewer and author is unknown to each other. This process guarantees the same chances for unknown and famous scientist and universities by removing the name on the submissions.

There are further versions of peer reviewing like open peer reviewing or additions like post-publication peer reviewing, but they are rarely been applied [2][3].

The crucial step for the quality of the peer review process is to find the best suited reviewer for each of the submitted papers. This person must fulfill the following two conditions:

- He should be an expert on the topic of the paper in order to give a qualified judgment on novelty, contribution and other aspects of the work presented.
- He must not be in any kind related to the author to guarantee a neutral statement about this paper, which means that there is no conflict.

After the conflicts of the reviewer are identified, he needs to be assigned to one or multiple papers. But before they can be assigned, an indicator is needed to measure which reviewer suits best to which paper. Most systems use a so called bidding process. In this case, the IPC member can indicate what papers he wants to review and in which areas he considers himself as an expert. This process is tedious for IPC members of conferences with hundreds of submissions. To address this problem, this paper presents an automatic approach by using the TF/IDF algorithm for matching the submitted papers with existing publications of the IPC member.

 ISBN: 978-1-61208-342-1

The next section will give an overview of other systems and techniques used in this domain. Section III and IV will introduce the TF/IDF and our implementation of PRIMA. The last two sections will address the results and the future work on PRIMA.

II. Related Work

There has already been some research on how to create a good matching between a reviewer and submission. Charlin et al. created a framework based on a machine learning techniques [4]. Dumais et al. examined Latent Semantic Indexing methods [5] for assigning reviewers to submissions. Hettich et al. and Basu et al. extracted with TF-IDF important words in submissions and mined the web for possible reviewers based on the extracted TF-IDF terms [6], [7]. In the context of submission paper to IPC matching further problems arise when there is a given amount of IPC members:

A. Conflict detection

Current systems use different approaches for conflict detection. For example, in EasyChair, one of the largest conference management systems [8], the IPC member manually specifies for which papers he has a conflict with the author [9].

Confious [10] uses an automatic approach in order to detect conflicts by comparing email suffixes or affiliation data. The problem there is that people do have more than one email address and they often do not use their institutional email address but rather an address from a large email service (e.g., Yahoo, GMail, GMX, etc.).

A more robust approach in finding these conflicts is implemented in SRMv2 [11] by queering the Digital Bibliography & Library Project (DBLP) [12] and checking if the IPC member and the author have a co-authorship. If there is a co-authorship found on the DBLP it indicates also a conflict of interests for further submissions [13].

B. Reviewer suitability

The reviewer of a paper must be an expert in the area of the paper. For this reason, the review assignment can not be done on a random basis. So the conference management system needs a method to rate how suitable a reviewer is for a submission. Many of the current systems use some kind of bidding mechanism to generate these values. These bidding systems can be separated in two classes:

- An IPC member manually bids on the areas of expertise. During the submission phase an author can classify his paper according to a predefined topic list. These topics can be special areas defined for a conference or a general classification scheme for example the ACM classification [14]. The IPC member receives the same list in order to define his own preferences in what fields he considers himself as an expert. An IPC member who is an expert in an area is a possible candidate for reviewing papers of this specific area.
- The IPC member manually bids directly on the papers. Based on the title and abstract of a submission the members can decide if they are qualified to review it or not.

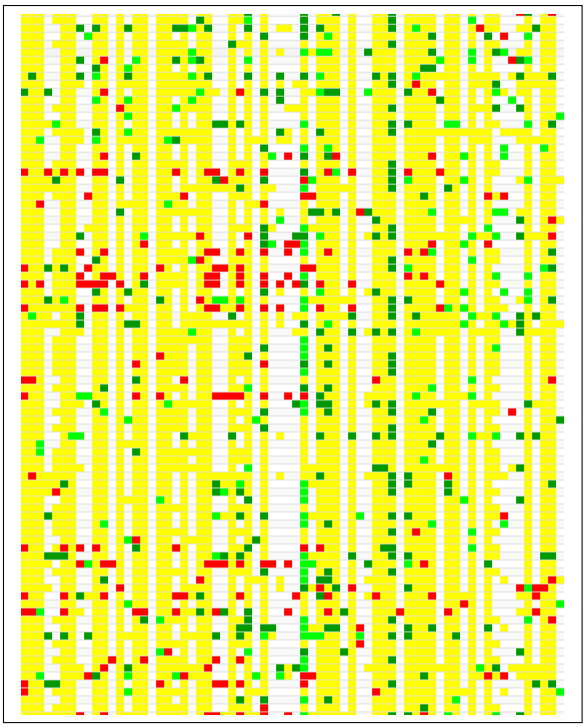


Figure 1: This figure shows a screen shot of the global bidding matrix. There it can be seen that the default value and the empty values take up most of the bidding values.

Larger conferences with hundreds of submissions sometimes combine these two options. As it is a considerable effort for an IPC member to read over hundreds of titles and abstracts, some systems let the users first fill out the topic list. Based on these data the systems generate a pre-ordered list of submissions for each IPC member. After that, the IPC members can read the title and abstracts of these submission, which fit best to their expertise profile. The resulting ratings can be used in the assignment process. This system however has two major drawbacks:

- Rodriguez et al. show in their paper ”Mapping the Bid Behavior of Conference Referees” [15] that human-driven referee bidding may not be the best solution for conference bidding due to referee fatigue. After reading several titles and abstracts an IPC member can have another decision basis. Furthermore, this bidding technique is susceptible to sloppy biddings due to curiosity or to unclear title and abstract of a paper. Using this bidding method, an IPC mehis area of expertise.
- The second issue about this system is that it doesn’t scale very well. An IPC member might be fine with reading some titles and abstracts. But if a conference has several hundreds submissions the effort for an IPC members is too large. Furthermore, it is not reasonable for all members to read all abstracts and have the same objectivity towards the last papers compared to the first. In addtion, if a reviewer reads only the papers in his expert area some papers which would fit to his expertise might be unnoticed.

 ISBN: 978-1-61208-342-1

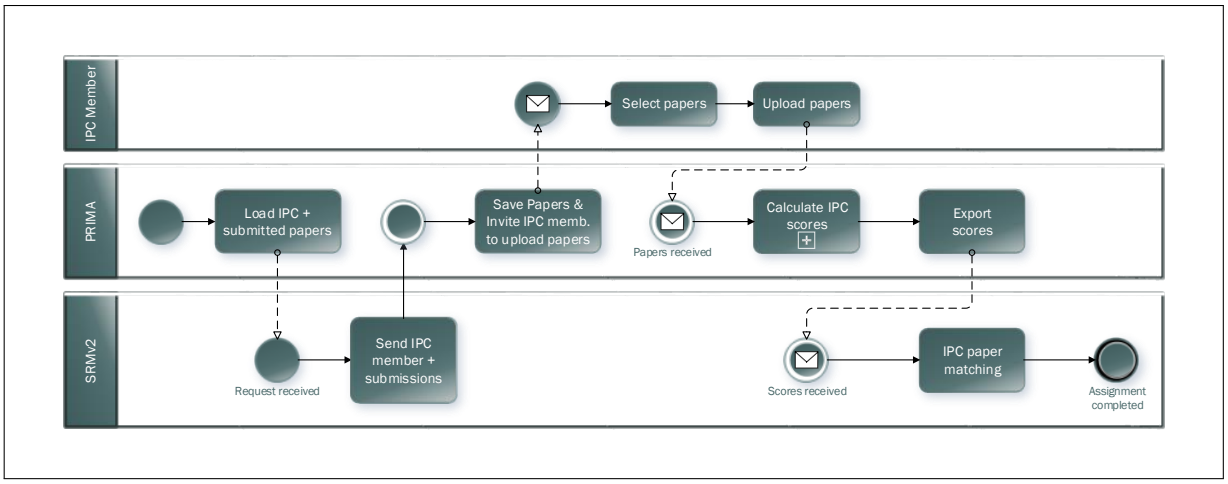


Figure 2: This figure shows the PRIMA workflow: how PRIMA receives the data and invites the IPC members to upload their papers. When all papers are received the calculation can start. After that the final scores can be exported again.

Figure 1 shows the bidding matrix for an exemplary conference. In this examples the IPC members (columns) specified for all papers (rows), whether they *want to review* (dark-green), *could review* (light-green), *have a conflict* (red) or are *not competent* (yellow). A large portion of the bidding matrix is either not filled up (white cells) or marked as *not competent* (yellow cells).

Our approach towards these problems is to automate the reviewer suitability rating by using the *Term Frequency Inverse Document Frequency* (TF/IDF) in order to categorize the submitted papers with respct to existing publications of the IPC members. These generated values can be used to refine and improve the values from the manual bidding process or even make the manual process obsolete. The huge advantage of this approach is, that it is possible to create a better IPC to paper distribution instead of a distribution of more or less randomly assigned reviewers.

III. TF/IDF

This section gives a small introduction to the mathematical basis of the TF/IDF. The term TF/IDF stands for *Term Frequency Inverse Document Frequency*. The TF/IDF algorithm can be separated into two parts. The first part is the *Term Frequency* part. As the name suggests it uses the frequency of terms in a document to classify the document. The second part of the algorithm is the *Inverse Document Frequency*. This means that the terms are weighted according to the occurrence in several documents. That is the more a term is used in different documents the less information it provides for classifying a document [16].

This paper will give an overview how the algorithm works. The algorithm itself is already a quite understood and researched topic in different areas like text categorization, text analysis, mining and information retrieval techniques.

In the term frequency calculation (see (1)) every term t in the document d is counted. For weighting the different terms in the document the logarithm is calculated. This is done because a term which occurs 10 times more than another term is not 10 times more meaningful.

$$tf(t,d) = log(1 + f(t,d)) \tag{1}$$

The inverse document frequency (see (2)) counts the occurrences of a term across all documents in a given document corpus. This is done by taking the logarithm of the quotient between the total number of documents $|D|$ and the amount of documents d containing the term t. Imagine a term which occurs in every document. This term is not useful for categorizing so it has to be penalized for being not important in the current global text corpus. Terms which occur in fewer documents receive a higher value with this formula.

$$idf(t,D) = log\frac{|D|}{|\{d \in D : t \in d\}|} \tag{2}$$

By multiplying the term frequency with the inverse document frequency the TF/IDF is received (see (3)). This value classifies a term in a document and its classification significance across all documents [17].

$$tfidf(t,d,D) = tf(t,d) \ \times \ idf(t,D) \tag{3}$$

All TF/IDF values of a document form a vector which classifies the document. By calculating the cosine similarity (see (4)) between two documents it is then possible to extract a similarity value [18]. By calculating the similarity between all submitted papers and previous papers of the IPC members we want to extract a matching value which enables matching submissions and IPC members together.

$$\cos(a,b) = \frac{a \cdot b}{\| \ a \ \| \| \ b \ \|} \tag{4}$$

There are many abbreviations in the TF/IDF algorithm, which offer different advantages and disadvantages. In the current version of our implementation the above described TF/IDF algorithm is used. Further research will show if another version or combination of algorithm yield to better results.

 ISBN: 978-1-61208-342-1

IV. PRIMA - Paper Rating and IPC Matching Tool

This section describes the prototype of the *Paper Rating and IPC Matching Tool* (PRIMA), which is a standalone extension to the SRMv2 conference management system. The workflow of the automatic score generation with PRIMA is shown in Figure 2.

In the first step, the PRIMA tool is initalized with the required data for the IF/IDF calculation: the submitted paper along with their metadata and information about the IPC members of the event. PRIMA uses the API of the SRMv2 framework [13] in order to fetch the required information. After the initalization, the IPC members are invited to upload their publications which fit best to the scope of the conference. The more papers a user uploads into the system the better the algorithm can find different matchings to the submissions of the conference.

One critical issue we found during the tests was that some IPC members received an overall good score on every paper. During the investigations we found out that some of the uploaded data were conference proceedings. From this files only the paper of the person was extracted and analysed as the whole report distorted the expertise of the user.

Then for all submitted papers of the conference and all uploaded publications of the IPC members, the paper scores are calculated. Then, these scores are transmitted into SRMv2 in order to support the pre-ordering for the bidding process and to support the assignment process.

Before the calculation itself starts some preprocessing steps are necessary to improve the TF/IDF result:

- For all uploaded publications, the raw text is extracted from the Portable Document Format (PDF) documents. This extracted text contains a large number of unnecessary information, which do not have an impact on the paper classification, for example numbers, special characters, code, urls, email addresses, punctuation, authors, addresses, IDs, etc. Future work on TF/IDF concentrates how to separate the text which is useful for the TF/IDF score generation from the overhead part which interferes with the generation [19].
- In the next step, stop words are removed. Stop words are words which occur often in a text but do not add any informational value to the text. Some examples of this stop words are: and, or, the, an, important, however, just and so on. All these words are necessary for the creation of sentences. But two texts do not relate strongly to each other just because they have a lot of “and” together [20].
- In the last step before the TF/IDF is applied all words have to be stemmed. Stemming reduces words to their common root. For example ”overview” and ”overviews” are not the same words in a computational matching, so the word overviews is reduced to overview. These two words will then match in the algorithm [21].
- The TF/IDF algorithm counts the words, normalizes and then they are weighted according to the occurrences in the other documents see Section III.
- After the TF/IDF calculation has been completed, each submission is compared against all papers of the IPC members with the cosine similarity.

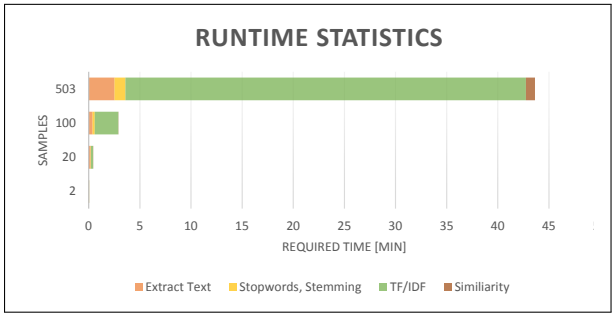


Figure 3: Runtime Statistic. This figure shows the amount of time each of the tasks take. It can be seen that the algorithm has an exponential growth.

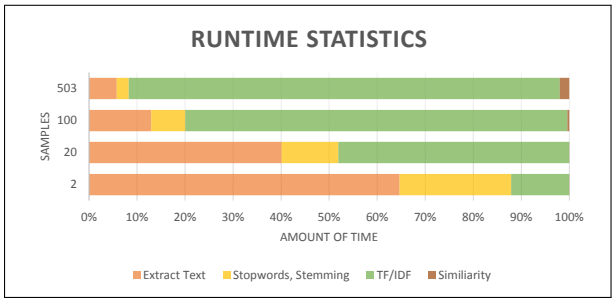


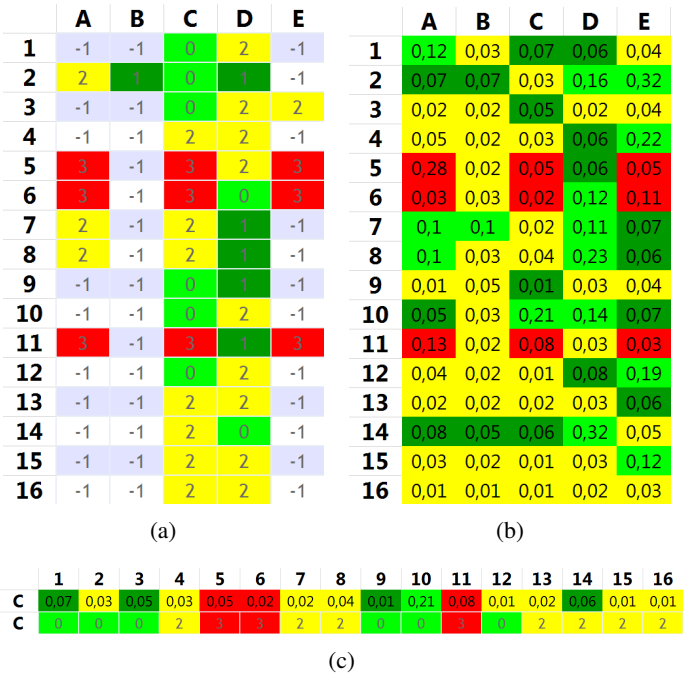
Figure 4: Runtime Statistics. This figure shows the amount of time each of the tasks take, split by the tasks and scaled to 100%.

If an IPC member has provided multiple publications, all of them are checked against a single submission paper. Currently, the average of the best five papers is saved. This is done to prevent statistical outliers. Furthermore, not all papers are taken into consideration as a person might upload a lot of papers belonging to different areas. In this case, every area on its own would have a lower average which falsifies the expertise area of a person. Further research will show if other values or a special algorithm should be used for creating a stronger statement about a submission and an IPC member.

V. Results

For testing the data the Eurographics 2014 was chosen. The papers, the submissions and the reviewers are anonymized and randomly reordered. Here are some statistics about the conference: There were about 70 programme committee members and about 290 submissions. Every IPC member entered their conflicts, defined areas of expertise to create a pre-filtering for the submissions and finally bidded on the paper. This final bidding matrix has 290 x 70 = 20300 entries (see Figure 1). For testing in PRIMA, about 300 papers of these IPC members were uploaded and together with the 290 submissions analyses through the TF/IDF algorithm.

 ISBN: 978-1-61208-342-1



	A	B	C	D	E
1	-1	-1	0	2	-1
2	2	1	0	1	-1
3	-1	-1	0	2	2
4	-1	-1	2	2	-1
5	3	-1	3	2	3
6	3	-1	3	0	3
7	2	-1	2	1	-1
8	2	-1	2	1	-1
9	-1	-1	0	1	-1
10	-1	-1	0	2	-1
11	3	-1	3	1	3
12	-1	-1	0	2	-1
13	-1	-1	2	2	-1
14	-1	-1	2	0	-1
15	-1	-1	2	2	-1
16	-1	-1	2	2	-1

(a)

	A	B	C	D	E
1	0,12	0,03	0,07	0,06	0,04
2	0,07	0,07	0,03	0,16	0,32
3	0,02	0,02	0,05	0,02	0,04
4	0,05	0,02	0,03	0,06	0,22
5	0,28	0,02	0,05	0,06	0,05
6	0,03	0,03	0,02	0,12	0,11
7	0,1	0,1	0,02	0,11	0,07
8	0,1	0,03	0,04	0,23	0,06
9	0,01	0,05	0,01	0,03	0,04
10	0,05	0,03	0,21	0,14	0,07
11	0,13	0,02	0,08	0,03	0,03
12	0,04	0,02	0,01	0,08	0,19
13	0,02	0,02	0,02	0,03	0,06
14	0,08	0,05	0,06	0,32	0,05
15	0,03	0,02	0,01	0,03	0,12
16	0,01	0,01	0,01	0,02	0,03

(b)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C	0,07	0,03	0,05	0,03	0,05	0,02	0,02	0,04	0,01	0,21	0,08	0,01	0,02	0,06	0,01	0,01
C	0	0	0	2	3	3	2	2	0	0	3	0	2	2	2	2

(c)

Figure 5: Figure (a) shows a small excerpt of the bidding matrix. Most reviewers set the values to the default not competent or did not submit any values at all. Figure (b) shows an excerpt of the PRIMA matrix with the same color encoding like the bidding matrix and thresholds at 0.05 and 0.1. Figure (c) shows the transposed bidding and calculated matrix of reviewer C for easier comparison.

Figure 5a shows a small excerpt of the bidding matrix. The rows represent five reviewers (a to e), the columns represent 16 submissions (1 - 16). The colors are encoded in the following way: Green means the IPC member submitted that he is able to review the paper. Whereas 0 (light green) means he wants to review the paper and 1 (dark green) means he could review the paper. 2 (yellow) indicates that the reviewer said he is not competent enough to review this paper. From -1 (white areas) we do not have any data as the IPC member did not bid on this paper. The red spots mark a conflict of the IPC member with the author of the submitted paper.

Figure 5b shows the same excerpt for the TF/IDF algorithm. The algorithm outputs a value between 0 and 1. Where 0 means no word overlap in both documents and 1 means every word in both papers appear at the same amount. Based on the global output we inked the output of the algorithm to give a similar appearance like the bidding matrix. The threshold of the values are 0.05 and 0.1. Everything below 0.5 is colored in yellow meaning a low correlation. Between 0.05 and 0.1 a medium correlation exists, which are marked in dark green. The high correlation ($>$ 0.1) are colored in light green. Additional the conflicts of the bidding are included in the results. To better compare these two tables the third figure shows the bidding matrix of the reviewer C to the calculated values.

A first observation is, that the left bidding figure shows that the provided data of the IPC is rather incomplete. This might happen because an IPC only checked the papers in his own area of expertise or because of a lack of time he was not able to read all 290 submission abstracts.

Another important observation is that some good matchings are conflicts (see at cell C11 of the left figure). This shows that the approach itself is heading in the right direction as it can be expected that a person who is an expert in an area also might have a project cooperation other experts in this field and therefore has a conflict of interests with this person.

Furthermore, it can be observed that most of the bidding match with the found generated values of PRIMA C1, C3, C9, C10 (Figure 5c). In addition, also the not competent column matches with the biddings C4, C7, C8, C13, C16. In this case, it should be said that the uploaded data of each person were taken only from previous Eurographics events and that the amount of uploaded data also differs. For example IPC member D has 18 uploaded papers and person B only five. For this reason person D is much better classified by the TF/IDF and therefore has a better matching than person B.

Strong differences between the bidding and the calculated classification, e.g., for person C the cells C2, C12, and C14, can have multiple reasons.

According to the TF/IDF the IPC member would be well suited as a reviewer, but he considered himself as not competent. This can have different reasons:

- The TF/IDF has analyzed an older paper of the person, but the expertise focus of the person has changed.
- The title and abstract from the bidding might have been misleading.
- The submission was overlooked by the IPC member and this submissions stayed on the default value which is *not competent*.

The first item will be addressed in further research in order to analyze if penalty value for older paper will improve the results.

But also cases where the rating from the TF/IDF shows a low score, but the persons claimed that he *wants to review* occur, for example in C2 and C12:

- Most likely the system does not have a current paper of the IPC member on this topic.
- The reviewer is interested in a paper and ”wants to review” it, but does not have the necessary knowledge to review it.

As stated before a large portion of the bidding matrix is not filled up (see Figure 1). One huge advantage of the process is when there is no value on the bidding matrix but the calculation found excellent matches on the algorithm. There it is possible to create a better reviewer-to-paper assignment instead of randomly distributing the submitted papers to the reviewers. For example in submission 4 the best matches are person D and E, for submitted paper 13 person E would be a good choice.

Figure 4 and Figure 3 show the runtime statistics of the PRIMA tool split into the four steps *extract text*, *stopwords,*

 ISBN: 978-1-61208-342-1

stemming, *TF/IDF*, and *similiarity*. It can be seen that for a small number of papers the extraction of the text and the stopwords removal and stemming takes up most of the time. If the number of papers increases, the more time the TF/IDF algorithm itself takes. The text extraction and stopwords removal and stemming can be precalculated and stored. However, this step takes less than 10% of the time during the full calculation using more than 500 papers. The TF/IDF itself cannot be precalculated as every further submission changes the weighting of each word in the calculation process. So it has to be calculated when all papers of the IPC members are available. The algorithm on 503 papers takes up about 45 minutes, Where 100 papers only take about 3 minutes and 2 and 20 papers are calculated in seconds.

VI. Conclusion & Future Work

In this paper, we presented the PRIMA tool, which automatically calculates a ranking between submitted papers and the available reviewers (IPC members). By using the TF/IDF for categorizing the submitted papers along the reviewers expertise, the workload of the reviewers and the chairs is reduced dramatically. TF/IDF itself is already a well researched topic in text categorization and information retrieval techniques.

One large problem in comparing various tools and their performance is the lack of a standardized benchmark for this task. All work so far, has used real data for evaluating and testing. Since they contain sensitive information, these datasets cannot become publicly available, which makes a direct comparison impossible.

For the upcoming Europgrahics conference it is planned to evaluate the scores by presenting submissions to the authors in descending order. Then the IPC member can concentrate on the title/abstracts which fit best to the topics of his own publications. The values that the PRIMA tool generates can also be used as suggestions for the reviewer during the bidding process. This way the member can skim over the values and check if they fit.

Currently the selection and upload of publications is done manually by the reviewers. Using citation portals like DBLP [12], Citeseer [22] and other sources, the selection and retrieval of the full-text version (e.g., when available through the Open Access [23] initiative) can be automated as well.

Another important point which might be improved is the text extraction itself. At the moment, the whole paper is used for the TF/IDF. And although the numbers, special characters, URLs, stopwords, etc., are removed there are still words which slip through which should not be used for the analysis. For example words like the author, the institution, figure explanations, headings, formulas and so on.

References

[1] Academia Publishing, “What is Peer Review?” 2014, [retrieved: 03, 2014]. [Online]. Available: http://academiapublishing.org

[2] R. M. Blank, “The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review,” American Economic Review, vol. 81, no. 5, December 1991, pp. 1041–67, [retrieved: 03, 2014]. [Online]. Available: http://ideas.repec.org/a/aea/aecrev/v81y1991i5p1041-67.html

[3] M. W. Consulting, “Peer review in scholarly journals: perspective of the scholarly community–an international study,” Author, Bristol, UK, 2008.

[4] L. Charlin and R. S. Zemel, “The toronto paper matching system: An automated paper-reviewer assignment system,” in ICML 2013 Workshop on Peer Reviewing and Publishing Models., Atlanta, Georgia, USA, Jun. 2013.

[5] S. T. Dumais and J. 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. ACM, 1992, pp. 233–244.

[6] C. Basu, H. Hirsh, W. W. Cohen, and C. Nevill-manning, “Recommending papers by mining the web,” in Proceedings of the IJCAI99 Workshop on Learning about Users, 1999, pp. 1–11.

[7] S. Hettich and M. J. Pazzani, “Mining for proposal reviewers: lessons learned at the national science foundation,” in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006, pp. 862–871.

[8] L. Parra, S. Sendra, S. Ficarelli, and J. Lloret, “Comparison of online platforms for the review process of conference papers,” in CONTENT 2013, The Fifth International Conference on Creative Content Technologies, 2013, pp. 16–22.

[9] Cool Press Ltd, “EasyChair conference system,” 2013, [retrieved: 03, 2014]. [Online]. Available: http://www.easychair.org/

[10] M. Papagelis and D. Plexousakis, “Conf!ous - The Conference Nous,” 2013, [retrieved: 03, 2014]. [Online]. Available: http://www.confious.com/

[11] C. Caldera, “Srm 2.0,” 2013, [retrieved: 03, 2014]. [Online]. Available: https://srmv2.eg.org/COMFy

[12] M. Ley et al., “DBLP Computer Science Bibliography,” 2013, [retrieved: 03, 2014]. [Online]. Available: http://www.informatik.uni-trier.de/ ley/db/

[13] C. Caldera, R. Berndt, and D. W. Fellner, “Comfy - A Conference Management Framework,” Information Services and Use, vol. 33, no. 2, 2013, pp. 119–128, [retrieved: 03, 2014]. [Online]. Available: http://dx.doi.org/10.3233/ISU-130697

[14] The Association for Computing Machinery, Inc., “Association for Computing Machinery,” 2013, [retrieved: 03, 2014]. [Online]. Available: http://www.acm.org/about/class/class/2012

[15] M. A. Rodriguez, J. Bollen, and H. V. D. Sompel, “Mapping the bid behavior of conference referees,” Journal of Informetrics, vol. 1, 2007, pp. 06–0749.

[16] J. Ramos, “Using TF-IDF to Determine Word Relevance in Document Queries,” Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855e, Tech. Rep., 2003.

[17] C. D. Manning, P. Raghavan, and H. Schtze. Cambridge University Press, 2008, [retrieved: 03, 2014]. [Online]. Available: http://dx.doi.org/10.1017/CBO9780511809071.007

[18] A. Huang, “Similarity Measures for Text Document Clustering,” in New Zealand Computer Science Research Student Conference, J. Holland, A. Nicholas, and D. Brignoli, Eds., Apr. 2008, pp. 49–56. [Online]. Available: http://nzcsrsc08.canterbury.ac.nz/site/digital-proceedings

[19] E. Rahm and H. H. Do, “Data cleaning: Problems and current approaches,” IEEE Data Eng. Bull., vol. 23, no. 4, 2000, pp. 3–13.

[20] C. Silva and B. Ribeiro, “The importance of stop word removal on recall values in text categorization,” in Neural Networks, 2003. Proceedings of the International Joint Conference on, vol. 3, 2003, pp. 1661–1666 vol.3.

[21] M. Porter, “An algorithm for suffix stripping,” Program: electronic library and information systems, vol. 14, no. 3, 1980, pp. 130–137.

[22] The Pennsylvania State University, “CiteSeer,” 2014, [retrieved: 03, 2014]. [Online]. Available: http://citeseerx.ist.psu.edu/

[23] Georg-August-Universitt Gttingen Niederschsische Staats- und Universittsbibliothek Gttingen, “Open Access,” 2013, [retrieved: 03, 2014]. [Online]. Available: http://open-access.net

 ISBN: 978-1-61208-342-1