Automating the Management of Scientific Conferences using Information Mining Techniques

Andreas Pesenhofer[1], Rudolf Mayer[2], Andreas Rauber[1,2]
[1] iSpaces Group,
eCommerce Competence Center – EC3
Donau-City-Straße 1, A–1220 Wien, Austria
andreas.pesenhofer@ec3.at
[2] Department of Software Technology and Interactive Systems,
Vienna University of Technology
Favoritenstraße 9–11/188, A–1040 Wien, Austria
{mayer, rauber}@ifs.tuwien.ac.at

Journal of Digital Information Management

ABSTRACT: *Conference management systems, which help organizers in carrying out tasks in the workflow, are widely used in the academic world. In this paper we focus on tasks where methods from the domain of information retrieval, information management and information organization can assist the organizer, the program committee members and the participants. We present a method for the creation of an improved review process by better matching the reviewers expertise with the paper topics, which can increase the quality of the conference. Furthermore the conference participants benefit from the better access to the wealth of information accumulated throughout a conference series. The conference organizers profit from the reduced workload because of the partial automating of tedious tasks, such as the review assignment, the compilation of the conference program and the creation of poster setup plans. We report on case studies from a small-sized (around 200 participants), a medium-sized (around 400 participants) as well as a large (more than 700 participants) conference in the computer science as well as the medical domains.*

Categories and Subject Descriptors
G.3 [Mathematics of Computing]: Probability and Statistics—Statistical computing; **H.3.3** [Information Storage and Retrieval]: Information Search and Retrieval—Clustering; **H.4** [Information Systems Applications]: Miscellaneous; **I.5.4** [Pattern Recognition]: Applications—Text processing

General Terms
Information mining, Information Orgnaization

Keywords: Automation, Assignment problem, Clustering, Self-organizing maps, Information visualization

Received 10 Aug. 2006; Revised and accepted 27 July 2007

1. Introduction

The main aim of scientific conferences is to make the dissemination of ideas possible and to make them visible to the public. Conferences are furthermore the optimal place to meet national and international scientists in the particular research field in question for exchanging ideas and for networking. Especially for young researchers (students), such events are very useful for learning how the research community works.

A report about the amount of organized meetings at the country and city level for the year 2005 was released by the International Congress & Convention Association (ICCA)[1]. These rankings cover meetings organized by international associations which take place on a regular basis, have more than 50 participants and rotate between a minimum of three countries. For the year 2005 the ICCA Data researchers have identified 5,315 events, a rise of 6.37% compared to 2004. These statistics are in-line with the rise of conferences and workshops announced over the DBWorld mailing list[2], where from 2000 to 2006 each year the continual increase in conferences lies between 12 and 33.78%. The growth of conferences inevitably tends to also result in the growth of unexperienced conference organizers (i.e. persons who have not organized a conference yet). The organization of a scientific conference is a challenging endeavor where a small error can have tremendous influence on the event. The IEEE, for example, provides a conferences organization manual[3] to reduce the risk. For the technical (scientific) part of the conference the use of web-based management systems (such as [2, 7, 8, 12–14]) is indispensable in handling the huge amount of submissions and reviews. These systems fulfill the basic requirements and drastically ease organization. Yet, there are still many tasks where methods from the domain of information management and information visualization can assist to further improve the quality of the scientific program as well as to reduce the workload of the organizers.

This paper describes the tasks in a conference management system where the use of information mining capabilities provides advanced methods to assist the organizer, the program committee member and the participants, extending the work presented in [9]. The focus of this paper is on the partially automated compilation of the scientific sessions and on the post-conference participation support for the 1st International Conference on Digital Information Management (ICDIM'06). The remainder of this paper is structured as follows. Section 2 gives an overview of related work. Section 3 describes the basic functionalities of a conference management system and four core tasks for further automatization will be tackled in Sections 4–7. Finally, we give a conclusion and present future work in Section 8.

2. Related Work

Conference management systems are web-based systems that assist the organizers carrying out tasks in the workflow of an academic conference. Such tasks are, for example, the collection of submissions, the handling of assigned papers that the Program Committee (PC) members have to review, the download of papers, the handling of reviewers' preferences and

[1] http://www.iccaworld.com

[2] http://dbms.uni-muenster.de/conferences/

[3] http://www.ieee.org/web/conferences/mom/

bidding, review progress tracking, web-based PC meetings, notification of acceptance/rejection and sending e-mails for notifications to authors or PCs. Once a bidding process has been performed, the assignment is handled as an optimization problem to allocate papers according to reviewer preferences while striving for equal load distribution. For the automatic assignment of reviewers to papers additional information from the authors concerning their interests is needed.

Dumais and Nielsen [3] used data given by 15 reviewers that consisted not only of the submitted abstracts and/ and/or interests, but also provided complete relevance assessments for the 117 submitted papers. Information retrieval principles and latent semantic indexing were used to generate the automatic assignments for each reviewer. This method achieved an improvement of 48% compared to the random assignment where on average four relevant documents out of the ten are selected.

Yarowsky and Florian [16] focused on the classification of every paper to exactly one of six conference committee members. They used 92 papers which were submitted to the ACL conference in electronic form and additionally requested committee members to provide representative papers so that a reviewer profile could be created. First a centroid for each reviewer and then a centroid for each committee as the sum of its reviewer centroids was computed. For each paper the cosine similarity was computed and compared with the committee centroids where the highest rank was the selection criterion. They concluded that the automatic methods could be as effective as human judges, especially in case where the judges may be less experienced.

In [10] the assignment of papers is done based on previous collected user ratings. The paper describes a method which provides an approximate solution to the problem without requiring each user to rate each item. The method relies on an interactive process where in each step (or ballot) the users have to rate a sample of items. Collaborative filtering is then performed to predict the missing ratings as well as their level of confidence. Performing a new ballot may improve the accuracy of the prediction. This algorithm tends to lead to a suboptimal solution if only a sub group of reviewers rates the ballot and if only one ballot round is performed.

In [12] the assignment is made based on the bids for special papers and on the the reviewers' expertise on the conference topics and the willingness to review papers on these topics. The reviewers may bid in several stages and the bids are accumulated. Graph theory is applied to carry out the assignment.

The most recent work in this domain was carried out by Aleman–Meza et al. [1] where they describe a semantic web application that detects conflict of interest relationships among potential reviewers and authors of scientific papers. The degree of conflict of interest between the reviewers and authors is calculated based on a populated ontology. As input they integrated entities from two social networks, namely 'knows' from a FOAF (Friend-of-a-Friend) social network and 'coauthor' from the underlying co-authorship network of the DBLP bibliography. This allows them to detect more potential conflict of interests than the simplified method that is implemented in [8].

3. Conference Management

3.1. User Roles and Tasks

In a conference management system users with different roles have to have access to specific tasks in a predefined time slot. An analysis of these roles and tasks is given in [4] and [8]. We have to distinguish between organizers, PC members or reviewers, authors, participants and persons visiting the web page. The program committee (PC) chair is in charge of the coordination and monitoring of the necessary tasks.

Such tasks include setup/customization, paper submission, conflict of interest detection, reviewer assignment, reviewing, paper selection, session creation, poster setup plans and conference participant support (c.f. Figure 1). In this paper we will concentrate on four tasks: the conflict of interest detection together with the automatic assignment of submissions to reviewer, the compilation of sessions as well as the creation of poster setup plans and the conference participant support as highlighted in Figure 1.

3.2. Tasks for Improved and Further Automatization

In this section we will focus on tasks where further automatization eases the work of PC members and the PC chairs. We more over try to identify means to assist conference participants both at and particularly after the conference in order to make the most of the wealth of information presented during the meeting and accumulated over the years in a conference series.

3.2.1 Task: Reviewer Assignment

The submission-to-reviewer assignment is done either automatically or manually by the PC chair, with an automatic assignment usually being followed by a manual adjustment. For the assignment the following constraints are taken into consideration:

- The submission topics should match with reviewer interests.
- A reviewer's bid for specific papers has to be taken into consideration.
- Reviewers should not get their own paper nor papers from a colleague from the same institution to review. A potential conflict of interest between the PC members and submissions has to be identified.
- Each PC member should get approximately the same amount of papers to review, so that they have an equal work load.

All these tasks rely on the input of the PC members. This can cause trouble if some of the PC members are reluctant or too busy to cooperate. It is not possible for the PC chair (administrator of the system) to make decisions for them, being limited to sending reminder mails and asking for their cooperation. For example, if only a small amount of the PC members choose the topics they are interested in and if only a few PC members register their bid, the assignment algorithm cannot work properly and produces suboptimal solutions, which have to be corrected manually by the PC chair.

In Section 4 we will focus on an automatic assignment of the submitted papers to the PC members based on their previous publications as a baseline for the manual bidding process. It

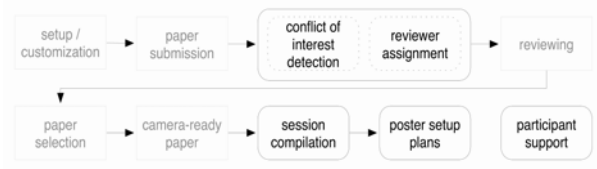


Figure 1. Processes in a conference management system

overcomes the above mentioned problems by using publicly available publications of the authors to create the PCs' profiles to identify potential interest/expertise matches as a baseline bid for those PC members who do not provide any specific bids.

3.2.2 Task: Session Compilation

After the selection of the papers has been completed, the program chairs have to find to an appropriate way to compile the scientific sessions. In the submission phase, the authors normally have to choose one or more research topics that are addressed in their papers. Sometimes it may also be required to select keywords from a pre-defined list that highlight the topical focus of the paper. If this information is available it can help the technical program committee in grouping accepted papers into sessions. Additionally, automatic clustering algorithms as described in Section 5 can be used.

3.2.3 Task: Poster Setup Plans

As part of most conferences, posters are presented in a special room or in the lounges of the conference venue. Usually there exists a pre-setup provided by the organizer where authors have to fix their posters. In this case the organizers have to figure out which posters fit best together when grouped by topic. Currently the PC chair has to align the posters manually. Mnemonic SOMs as described in Section 6 can be used to automatically create an assignment.

3.2.4 Task: Participant Support

The conference program should be kept up to date in the Web and the proceedings should be searchable either publicly or limited to registered conference participants via dedicated logins. Participants may be interested if they have missed interesting sessions. Mnemonic SOMs and SOMs in combination with the participant's interests give the participants new insight into the huge amount of information presented during the conference as well as helping them to prepare their schedule before attending large events. We will address that in more detail in Section 7.

3.3. Case Studies

We report on case studies from three conferences, the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)[4], the 1st International Conference on Digital Information Management (ICDIM'06)[5] and the European Congress of Radiology 2004 (ECR 2004)[6]. The ECDL is the major European conference on digital libraries and associated technical, practical and social issues in this field. It can be classified as a medium-sized conference with around 100 to 200 submissions, around 80 to 90 program committee members and around 350 to 450 participants. Due to the fact that the ICDIM took place for the first time, a manageable amount of participants (up to 200) attended the conference and, in addition to the scientific sessions, offered a number of workshops and tutorial sessions. The ECR is a large-sized conference series with more than 2,000 scientific paper submissions, taking place every year in Vienna. It is the largest radiological meeting in Europe attracting more than 15,000 participants from over 90 countries. WEBGES[7], who provide the soft- and hardware for the ECR, also provided us with the relevant data relating to paper and poster submissions.

[4] http://www.ecdl2005.org
[5] http://www.icdim.org/icdim2006/
[6] http://www.ecr.org
[7] http://www.webges.com

The data has to be transformed into a numerical representation understandable and processable by computer systems. Therefore, we indexed the documents based on the the well known bag-of-words approach with Lucene[8] using a tf×idf weighting scheme [11], which is based on the term frequency (tf) in the given document and the inverse document frequency (idf) of the term in the whole collection. We applied several pre-processing steps to remove all punctuation marks and special characters.

An English stop word list was used to remove high frequent terms, additionally term reduction methods based on document frequency and term length were applied and finally regular expressions were used to remove numbers, dates, email addresses and URLs. For the remaining terms we calculated the tf×idf values, which were normalized to unit vector length, so that the documents' length has no influence on the weight.

3.3.1 ECDL Corpora

For the ECDL we have to distinguish between three corpora:

ECDL A: Is made out of 723 automatically retrieved publications from PC member's home pages and 125 submissions. Applying term reduction methods as mentioned above resulting in a vector with 8,767 unique terms.

ECDL B: Consists of the accepted poster submissions, 30 different posters in the English language. After preprocessing and term reduction we obtained a feature space of 569 different terms.

ECDL C: Is composed of the accepted paper and poster submissions, totaling to 71 documents. Applying the same mechanisms as before we obtained a vector of 5,654 different terms. In all three cases no stemming was applied.

3.3.2 ICDIM Corpus

The ICDIM corpus consists of the scientific papers from the conference proceedings, excluding the papers from the workshops. In total, this collection consists of 85 papers, each of which belonging to exactly one research topic. We applied the categorization scheme as it was used in the proceedings resulting in 15 topics (c.f. Figure 7). Again we applied the preprocessing steps and term reduction steps. In the end we obtained a vector of 3,020 different terms. Note that for one document (p131) it was not possible to extract the textual content from the provided documents.

3.3.3 ECR Corpus

This corpus consists of the abstracts of the ECR from the year 2004. All together there are 943 English documents which were presented during the scientific sessions of the congress and which each belong to one of the 15 different session topics (c.f. Figure 6). Every document is assigned to exactly one topic. After the preprocessing, the corpus consisted of 3,842 unique terms.

Additionally, we received the radio frequency identification (RFID) logs that were collected during the conference. At the registration every participant received a badge with a unique RFID tag. The entrances to the halls of the conference location were guarded with RFID gates, so that the organizer could track access to a session. The collected attendance information is used in the medical domain for the monitoring and issuing of continuous education certificates. They serve to build an anonymous participant profile for our experiments.

4. Profile based Reviewer Assignment

A good paper-to-reviewer assignment is based on the cooperation of the PC member (reviewer). They have to choose

[8] http://lucene.apache.org/

from a list of relevant topics which they are interested in and furthermore they have to bid for special papers by skimming through the abstracts. Most of the PC members neither bid nor choose their interests so that standard algorithms fail in computing a proper assignment. This is particularly due to the fact that a bidding process for 200 or more papers is a notoriously time consuming task. Our solution overcomes this problem, because the interest of the reviewer is defined based on his or her publications that are available on the Internet.

4.1. Profile Generation

We use the name of the PC members to formulate the search query, which is subsequently sent to two search engines which provide scientific papers, namely CiteSeer.IST[9] and GoogleScholar[10]. From the returned search result pages the URLs linking to the publications were extracted. Using the 87 PC members from the ECDL 2005 conference resulted in 4,369 retrieved URLs. In the next step we downloaded these documents discarding all non PDF documents. Additionally we used simple heuristics based on author name and document structure to verify that we are dealing with publications from PC members. As a result we obtained the 723 potential publications. For ten PC members no publications have been automatically retrieved. Those reviewers can themselves upload papers in the system to generate their profile.

4.2. Conflict Of Interest Detection

The potential conflict of interest detection (COI) was performed based on (1) the occurrence of the last name of a program committee member in the authors line of a submission and (2) the existence of parts from the PC members email domain

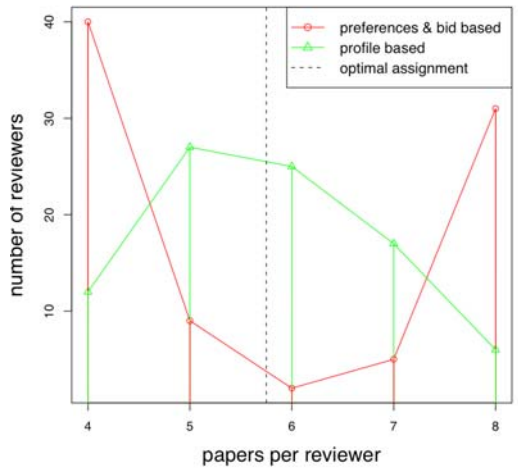


Figure 2. Distribution of the review workload

in the submissions author field. Using these two methods allowed us to identify 46 potential conflict of interest for the PC members for the ECDL 2005 data set.

We compared our results with the COI that the PC members registered during the biding phase of the ECDL conference.

Here in only 24 cases a COI was registered. A detailed comparison of the two lists reveals the following:

1. More than the half (57.69%) of the reviewers who should have registered a COI, did not bother to enter one. For this group of people we automatically inserted the COI.

2. A potential COI was detected by the system but not registered by the reviewer, who in principal did register 50% of the COI. As reasons we identified that the COI was not considered in spite of being from the same institution, because of a lack of close cooperation inside the institution and that the paper was overlooked due to the large list of papers. A solution would be to have a system that detects a potential COI and presents it to the reviewer to confirm it.

3. In seven cases the COI was registered by the reviewer, but not detected by our current system. In these cases coauthorship analysis would have to be included (e.g. DBLP11) and for areas that are not covered by a specific digital library of papers a web-based search has to be performed.

4.3. Reviewer Assignment

Before we can calculate the assignment, we have to detect which submissions match with the interests of which PC member. Therefore, we computed the Euclidian distance between every submission and publication based on the full-text indexed feature vector. A distance of 0 means that the two compared documents are identical, and the higher the value the more different they are. A PC member has normally more than one publication in his or her profile, so we keep only the smallest distance from all of his or her documents to a specific submission. This results in one distance value per submission, which were subsequently sorted from the smallest to the largest distance. The first ten received a rate level of 4 which correspond to a bid of 'eager' to review, the next ten were rated with 'interesting' (3) and the remaining received the level 1 ('better not'). For the ten cases where no publications could be found automatically, and therefore no distances to the submission existed we used 2 ('indifferent') as default rating. If a COI in the relation was detected, a rate level of 0 ('conflict of interest') was inserted into the data base.

As baseline for our evaluation we use the automatic assignment that was calculated on the ECDL 2005 PC member preferences and their bids. To make our system comparable with the baseline we set up an identical system without the bids and the paper topic interest of the PC members. We inserted the profile based bids into the system. These pre-calculated values serve as a basis for the bidding process that may be optimized by the PC members. In our experimental setting no further modification was made.

Figure 2 summarizes the workload distribution of the PC members using the assignment model based on preferences and bids compared to the results that where obtained with the profile-based assignment. In both cases we have 500 reviews that have to be assigned to the 87 PC members, the optimal amount of assigned papers per PC member would have been 5.75. In the first case, the preferences & bid-based model, 40 reviewers get four papers to review and 31 reviewers get the maximum amount of papers (eight) assigned. Only 16 reviewers get six to seven papers assigned. In our system, the profile-based one, only six reviewers have a workload of eight papers and twelve PC member have only four papers to review. Most of the PC members (27) got five papers, followed by 25 that got six and 17 that got seven papers assigned. In this case many more PC members are allocated around the mean of 5.75 resulting in a more equal distribution than in the first case.

5. Session Compiling

When the final versions of the papers are uploaded via the submission system, the program committee chair can start to shape the conference program. A fixed number of time slots for

[9] http://citeseer.ist.psu.edu/

[10] http://scholar.google.com/

[11] http://www.informatik.uni-trier.de/Üley/db/

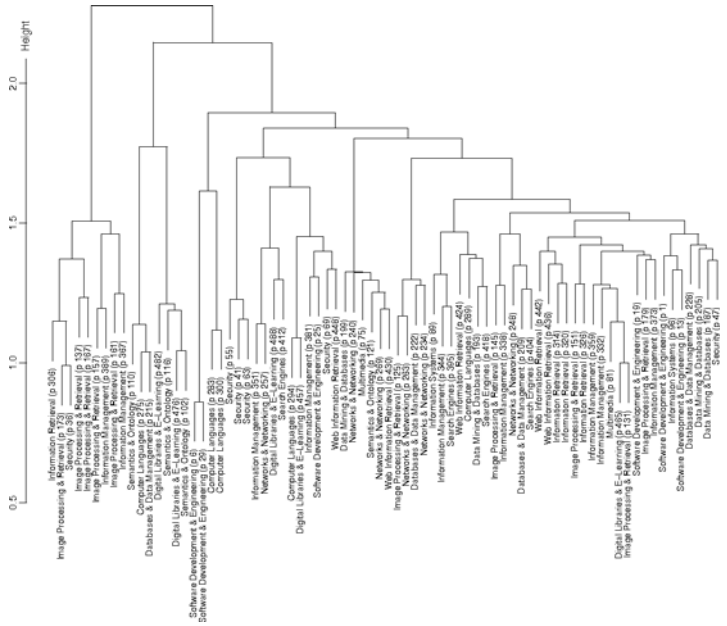


Figure 3. Cluster dendrogram of the ICDIM 2006 papers

scientific presentations is available to populate. The PC chair has to find an appropriate way to group thematically related papers together. The sessions can either be in parallel - especially when a huge amount of papers should be presented during the conference - or in sequential order with no overlaps. This is traditionally done based on the histogram of submissions per research topic. Note that the topic has been chosen by the authors. Sometimes, especially at huge conferences, the organizers also ask the authors to choose significant keywords so that the organizer can group related papers more easily. We propose to cluster the final papers with a hierarchical clustering algorithm to get a dendogram as shown in Figure 3. The Ward's linkage method [15], a minimum variance method, which aims to find compact and spherical clusters, was used. The hierarchical clustering approach produces an ordering of the documents, that may be helpful for the creation of the sessions. Documents that are further down in the tree structure are more equal to the linked document than documents that are located higher in the structure. As label for the documents we use the session names and the page number of the documents in the proceedings.

In the case of compiling sessions the PC chair may have a look on the graph and imagine horizontal lines at different height levels. A line at the level of two would result in three cluster where the first one consists of nine documents, the second of seven and the third pools the remaining documents. For the third cluster a line at the level of 1.7 might be drawn to further split up the cluster into sub clusters. We analyze the first cluster with the nine documents more briefly, to verify if our method generates comparable results with what was done by the organizers of the ICDIM. This cluster deals mostly with topics from 'information retrieval' and 'image processing and retrieval'. Three Documents from other topics are also found in this cluster. One of them (p36) comes from 'Security' and the other two (p389 and p367) are from 'Information Management'. The paper p36 entitled 'A RobustWavelet Based DigitalWatermarking Scheme Using Chaotic Mixing' describes a method to digital watermark digital images; the paper p389 entitled 'Word-wise Script Identification from Bilingual Documents Based on Morphological Reconstruction' deals with OCR from images in combination with k-nearest neighbor algorithm and the paper p367 deals with 'Recovery of Digital Information Using Bacterial Foraging Optimization Based Nonlinear Channel Equalizers'. The first two papers deal more or less with 'image processing' but the third one dose not fit that well into this cluster. Our method assists the organizers in compiling the sessions to a great extent and provides a good basis, however the final arrangement has to be done by the organizers.

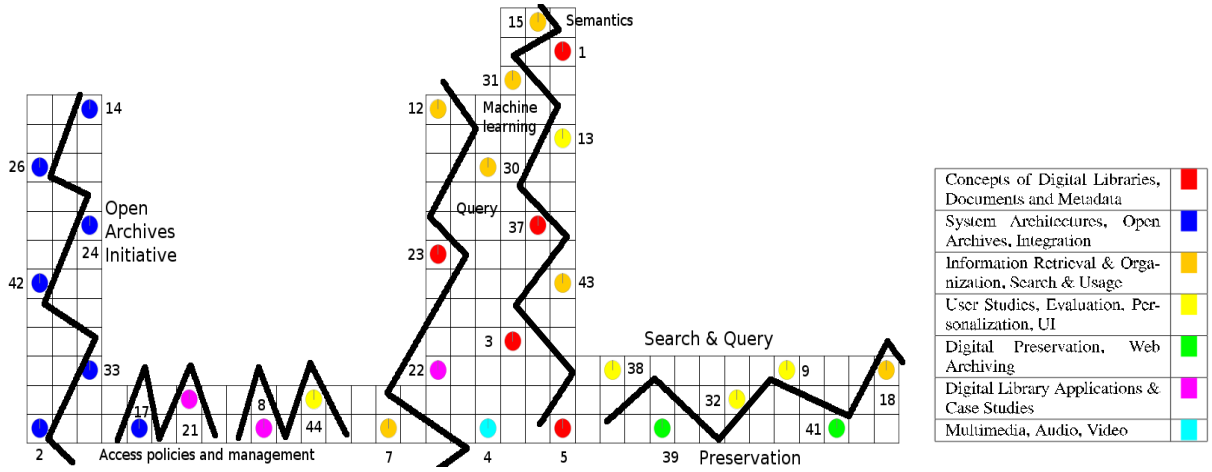
6. Poster Setup Plans

When the setup of the poster locations is defined by the conference organizers it can be done in one of several different ways.

For example, the setup may be organized in a completely random way or sorted alphabetically by author names or submission titles. It may be desirable, though, to organize the submissions by their content - that way, conference participants can easily find the areas with posters about topics they are interested in. Organization by content may be done using manually assigned category labels coming either from the authors themselves during submission, or from the PC. However, such a categorization may in many cases not be available at all, available only for some parts of the submissions or of poor or varying quality. Then, as an alternative, unsupervised clustering algorithms based only on the submission contents may be utilized to determine a poster setup.Independent of the exact setup, the conference participants should also be provided with a map of the venue, indicating poster locations and topic areas, in order to assist them in locating the posters they are interested in.

Both unsupervised clustering and generating a map of the poster setup can be achieved using for example the Self-Organizing Map (SOM) [5]. The SOM is a neural-network model that provides a mapping from a high-dimensional input space to a lower dimensional output space. In this mapping, the SOM preserves the topology of the input space, i.e. input patterns that are located close to each other in the input space will also be located closely in the output space, while dissimilar patterns will be mapped on to opposite map regions. In many applications, the output space is made of a two-dimensional, rectangular map. This representation allows for an easier interpretation of the complex structure of the input patterns by the user, due to its analogy to two-dimensional geographic maps.

Another advantage of using the SOM is that it generates a clustering that preserves transitions between clusters -



(a) (b)

Figure 4. Poster alignment for the ECDL 2005 conference

documents that would belong to two different clusters will be mapped on the border in between those clusters. In our application, the input space will be formed by a vector-space representation of the poster submissions, as described in Section 3.3, while the output space will be the map of the poster session area.

As in many cases, the area for the poster session may not be of rectangular shape, therefore we use a modification to the original SOM algorithm, the Mnemonic SOM, as presented in [6]. In the Mnemonic SOM, the output space is twodimensional, but can take any arbitrary shape. They can be easily generated from a black and white image representing the desired shape, for example the poster presentation area.

We have applied this method for arranging the poster setup during the ECDL 2005. Figure 4(b) gives an overview of the topics of the submitted posters to this conference. The category assignment was given by the authors on submission. Figure 4(a) shows the generated mapping, where the output space was made of a grid with the size of 35x15, with 182 units within the map shape. It is based on the layout of the conference poster area. Black lines show the setup of the poster boards, and numbers indicate the unique ID assigned to each poster on submission. The labels on the map (e.g. Query, Machine Learning, Preservation) have been added manually after inspecting the content of the documents grouped together in this region.

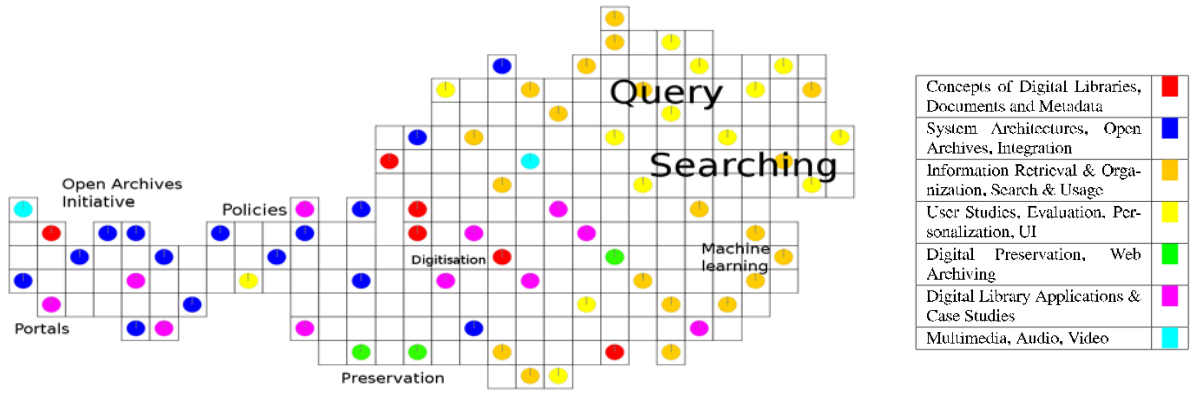
We can observe that thematically similar posters get arranged close to each other, for example in the top-left we can find posters dealing with the 'Open Archives Initiative Protocol'. The poster arrangement does not necessarily follow the manual categorization, but arranges them by content.

The given data set contains a lot of different, sometimes rather small clusters. This is due to the small size of the data set (30 accepted posters), and the very heterogeneous topics they discuss. However, the quality of the generated mapping is good. Using the method described above can help the conference organizer both in saving time on the poster setup and in achieving a better thematically grouped setup.

7. Participant Support

The method of the SOM, described in Section 6, can also be well utilized for supporting the participant during and after the conference.

One application is to provide an advanced interface to the proceedings of the conference, in addition to traditional key-word based searching or manually created indices. We again generate representations of all the presentations at the conference via a vector-space representation of the abstracts, and map the documents on a SOM. Figure 5 gives an example from the ECDL 2005 conference in Vienna (cf. Section 6), where we use a map in the shape of Austria as a mnemonic hint for the participants. The submissions, including both papers and posters, have been grouped automatically according to their topic by the SOM algorithm. The colored pie-charts visualize the distribution of the manually assigned categories of the documents. As in Figure 4(a), the labels have been placed manually after analyzing the content of the documents on the map.



(a) (b)

Figure 5. Map of the submissions to ECDL 2005

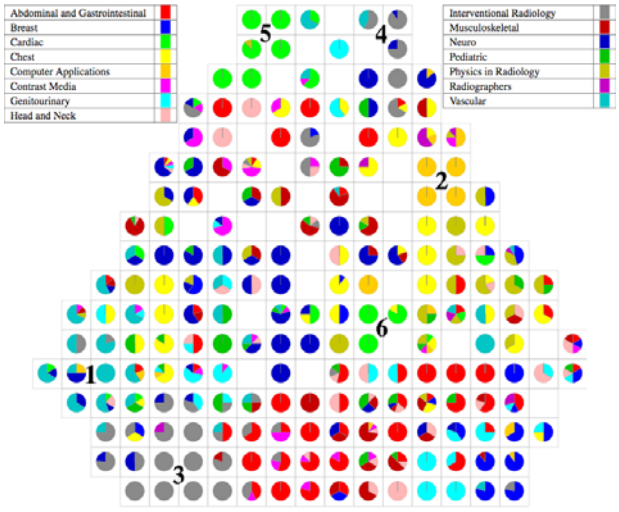


Figure 6. Scientific submissions to the ECR 2004 mapped on the ACV logo

The scientific abstracts of the ECR 2004 were also pre-processed as described in Section 3.3.3 and mapped onto a SOM, this time following the shape of the logo of the Austria Center Vienna (ACV), the location where the conference takes place every year. The contour of the logo also represents the basic shape of the ACV building. Due to the fact that the conference takes place every year at this venue, the shape serves as a mnemonic hint for the participants.

Figure 6 illustrates how this content based mapping of the shape of the ACV is done by the SOM algorithm. The legend shows the category names and colors of the ECR 2004, so that the evaluation of the map with the colored pie-charts can be done in an appropriate way. In the far left corner papers dealing with 'Vascular' (magenta; mark 1) are arranged together. The papers dealing with 'Computer Applications' (orange; mark 2) have their cluster on the right hand side. Papers dealing with 'Interventional Radiology' (grey) are split up into two clusters, where the first one (mark 3) deals with embolization and the second one (mark 4) deals with different kinds of stents. In the neighboring cluster (mark 5) the papers also deal with stents, in particular coronary artery stents, chest pain and thrombus detection belonging to the class 'Cardiac' (light green). In the second 'Cardiac' region (mark 6) the documents deal with ventricles, myocardial infarction and myocardial scars.

As a second example of this technique we have taken the papers of the ICDIM and created a representation as shown in Figure 7. As visual metaphor we used the 'Vidhana Soudha', an impressive building in the center of Bangalore containing the seat of the state legislature of Karnataka. On the top of the building documents with the main topic 'Information Retrieval' are grouped together. In the middle cupola they focus on 'Image Processing and Retrieval', further down the topics 'Web Information Retrieval' and 'Computer Languages' are relevant. In the upper right corner of this area (on top of the right spire), one single document from the group 'Multimedia' is mapped. As this publication, however, deals with topic maps for spatial-temporal multimedia blogs and their visualization, it fits into this area.

The version of this paper as found in the proceedings of the ICDIM conference is located in the flower bed in front of the entrance (marked with a diagonal star). The paper is surrounded by five units (squares), of which only the one in to the upper left is populated. This means that our paper is the most similar to the publication entitled 'AnT&CoW: Share, Classify and Elaborate Documents by means of Annotation'. This publication was also assigned to the 'Information Management' topic by the program committee chairs or the ICDIM.

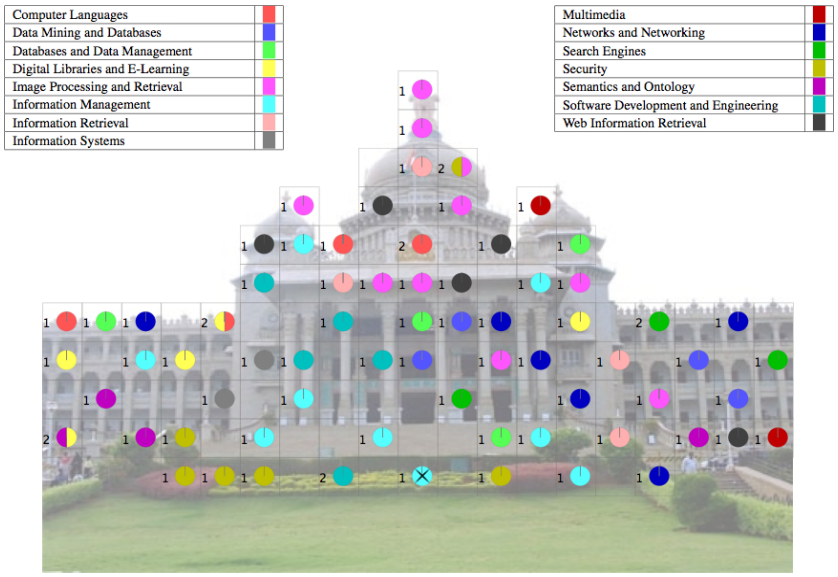
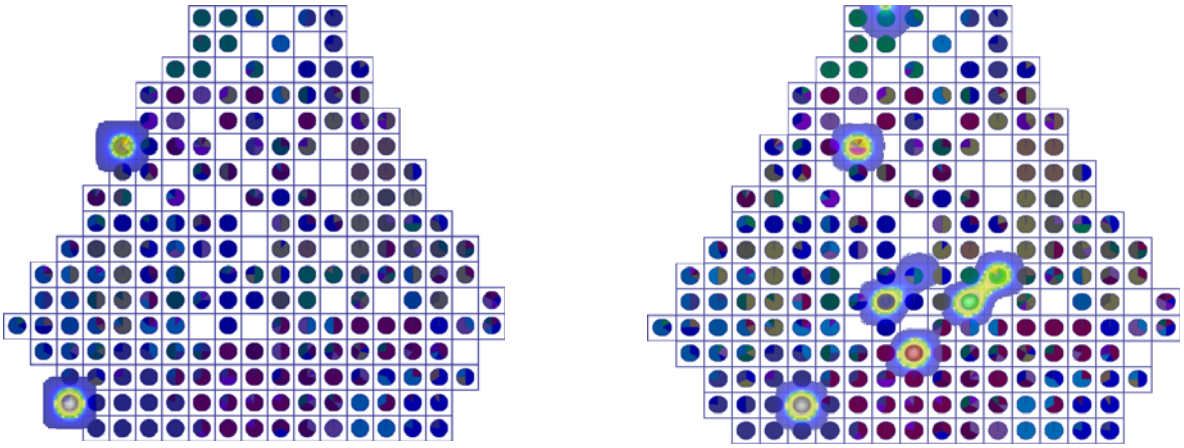


Figure 7. Scientific papers of the ICDIM 2006 mapped on the Vidhana Soudha



(a) Participant A

(b) Participant B

Figure 8. Fingerprints of ECR participants

In Figure 8 we used the attendance information of an ECR participant (RFID logs) to create personalized fingerprints. We identified the locations of these abstracts, which were presented in sessions that the participant attended, and created a hit histogram. The more focused a participant is, the more concentrated the histogram appears on the map. Participants can immediately see where their interests are located on the map. Documents of sessions that the participant visited at the conference are located in the highlighted regions, together with other documents that are similar in content. Thus, the surrounding regions may provide more insight into what others have presented at the conference. Figure 8(a) shows a participant who visited the two sessions 'Interventional Radiology' (lower left side) and 'Neuro' (top left side).

For the second participant (Figure 8(b)) six regions are highlighted. The two sessions entitled 'Myocardial viability and wallmotion' and 'Evaluation of cardiac function' both are part of the 'Cardiac' topic. They are located next to each other, forming a larger cluster on the right. Two documents from the last session are mapped to the top left of the map, where a second 'Cardiac' cluster can also be identified. The session dealing with 'Molecular Imaging' papers can be found half way down to the 'Neuro' region. Going down and a little bit to the right we come to the 'Musculoskeletal' session and going diagonal to the left we end up in the furthermost left spot, described as the 'Interventional Radiology' section. This participant attended sessions with five different topics, which can be seen by the fingerprint. For this participant it may be interesting to have a look at documents located in between the three clusters that are located close to each other. These might be documents that are interesting for her or him.

In the case of the ECR, the personalized fingerprint can be added to the profile of the participant. The accepted scientific papers of the upcoming conference can be trained as a mnemonic SOM in the shape of the ACV. Using the stored fingerprint allows the participants to mark their interests on the actual conference map and helps them to decide which of the sessions to visit.

8. Conclusion

In this paper we presented different information mining methods which can be used to improve scientific conference management systems. All the methods are content based; resources like the Internet are used for gathering additional information or access logs, when applicable. We showed that organizers, reviewers and participants of small, medium and large-sized conferences benefit from our proposed methods. To ease the task of paper to reviewer assignment for the organizer, and to subsequently reach a better quality of reviews we generate reviewer profiles, describing their review expertise, to identify the most fitting submitted papers. Then, we combined these results with scheduling algorithms to reach an evenly distributed assignment over all reviewers. Unsupervised clustering algorithms can identify similarities among the paper submission. We employed hierarchical clustering techniques to automatically propose a compilation of the papers into scientific sessions. Further, we assist the organizers in creating setup plans for poster sessions. We utilize an unsupervised neural network model, namely the Self-Organizing Map, to generate a two-dimensional map of the poster submissions. Using a variant of the SOM, the Mnemonic SOM, we can generate maps that are not rectangular, but can take the shape of the location of the poster session. In this map, topically similar posters will be arranged close to each other, thus resulting in a arrangement which is convenient for the visitors interested in specific topics. Using the SOM to generate a map of all the submissions, we further provide the registered participants better access to the scientific papers and also help them to decide which sessions they should visit at the next conference. We demonstrated our methods on case studies from three conferences of different size.

Future work includes the creation of a 3D virtual environment for presenting the content of a conference in an interactive way. The Mnemonic SOM will be used for the arrangement of the documents in the space.

9. Acknowledgments

This work was partially funded by the Austrian Federal Ministry of Economics and Labour under the k-ind research program and the MUSCLE Network of Excellence (project reference: 507752).

References

[1] Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Ding,L., Kolari, P., Sheth, A.P., Arpinar, I.B., Joshi, A.,Finin, T (2006). Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection. *In:* Proc. of the 15th International World Wide Web Conference, p. 407–416, Edinburgh, Scotland.

[2] ConfMan (Conference Manager). [Online]. Available: http://www.ifi.uio.no/confman/ABOUT-ConfMan/

[3] Dumais, S. T., Nielsen, J (1992). Automating the Assignment of Submitted Manuscripts to Reviewers. *In:* Research and Development in Information Retrieval, p. 233–244.

[4] Halvorsen, P., Lund, K., Goebel, V., Plagemann., T, Preuss, T, K¨onig, H (1998) Architecture, implementation, and evaluation of ConfMan: Integrated WWW and DBS Support for Conference Organization. Technical report, I-1998.016-R, UniK, University of Oslo, Norway.

[5] Kohonen, T (1995). Self-organizing Maps, volume 30 of Springer Series in Information Sciences. Springer Verlag, Berlin, Heidelberg.

[6] Mayer, R., Merkl, D., Rauber, A (2005). Mnemonic SOMs: Recognizable Shapes for Self-Organizing Maps. In Proc. of the 15th Workshop on Self-Organizing Maps (WSOM'05), pages 131–138, Paris, France.

[7] OpenConf Conference Management System. [Online]. Available: http://www.openconf.org/

[8] Papagelis, M., Plexousakis, D., Nikolaou, P. N (2005). CONFIOUS: Managing the Electronic Submission and Reviewing Process of Scientific Conferences. *In*: Proc. of the 6th International Conference on Web Information Systems Engineering (Industrial Track), New York, USA.

[9] Pesenhofer, A., Mayer, R., Rauber, A (2005). Improving Scientific Conferences by Enhancing Conference Management Systems with Information Mining Capabilities. *In:* Proc. of the 1st International Conference on Digital Information Management (ICDIM 2006), p. 359–366, Bangalore, India.

[10] Rigaux, P (2004). An iterative rating method: application to web-based conference management, *In*: Proc. of the 2004 ACM Symposium on Applied Computing, New York, NY, USA, 2004. ACM Press.

[11] Salton, G., Buckley, C (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24 (5) 513–523.

[12] The CyberChair Software. [Online]. Available: http://cyberchair.org/

[13] The Microsoft Conference Management Toolkit. [Online]. Available: http://msrcmt.research.microsoft.com/cmt/

[14] The MyReview System. [Online]. Available: http://myreview.lri.fr/

[15] Ward, J. H (1963). Hierarchical grouping to optimize an objective function. Journal of American Statistical Association, 58 (301).

[16] Yarowsky, D., Florian, R (1999). Taking the load off the conference chairs: towards a digital paper-routing assistant. *In:* Proc. of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very-Large Corpora., p. 220–230.

Andreas Pesenhofer is a researcher at the EC3 - E-Commerce Competence Center and member of the iSpaces research group. He received his MSc in social and economic sciences in 2004 from theVienna University of Technology. His main research interests include, but are not limited to, informationretrieval, information visualization, text mining, text classification and machine learning.

Rudolf Mayer is a researcher at the Department of Software Technology and Interactive Systems (IFS) at the Vienna University of Technology (TU-Wien). He received his MSc in social and economic sciences in 2004 from the Vienna University of Technology. His main research interests include, but are not limited to, information retrieval, information visualization, text and music mining, and machine learning, specifically neural networks and self-organising maps.

Andreas Rauber is Associate Professor at the Department of Software Technology and Interactive Systems (IFS) at the Vienna University of Technology (TU-Wien). He additionally is head of the iSpaces research group at the eCommerce Competence Center (EC3). He received his MSc and PhD in Computer Science from the Vienna University of Technology in 1997 and 2000, respectively. In 2001 he joined the National Research Council of Italy (CNR) in Pisa as an ERCIM Research Fellow, followed by an ERCIM Research position at the French National Institute for Research in Computer Science and Control (INRIA), at Rocquencourt, France, in 2002. His research interests cover the broad scope of digital libraries, including specifically digital preservation, text and music information retrieval and organization, information visualization, as well as data analysis and neural computation.