

## THE ISMIR CLOUD: A DECADE OF ISMIR CONFERENCES AT YOUR FINGERTIPS

Maarten Grachten   Markus Schedl   Tim Pohle   Gerhard Widmer

Department of Computational Perception  
Johannes Kepler University  
Linz, Austria

music@jku.at  
<http://www.cp.jku.at>

### ABSTRACT

In this paper, we analyze the proceedings of the past *International Symposia on Music Information Retrieval* (ISMIR). We extract meaningful term sets from the accepted submissions and apply term weighting and Web-based filtering techniques to distill information about the topics covered by the papers. This enables us to visualize and interpret the change of hot ISMIR topics in the course of time. Furthermore, the performed analysis allows for assessing the cumulative ISMIR proceedings by semantic content (rather than by literal text search). To illustrate this, we introduce two prototype applications that are publicly accessible online<sup>1</sup>. The first allows the user to search for ISMIR publications by selecting subsets of ISMIR topics. The second provides interactive visual access to the joint content of ISMIR publications in the form of a *tag cloud* – the *ISMIR Cloud*.

### 1. INTRODUCTION AND MOTIVATION

Music information retrieval and extraction has been a fast growing field of research during the past decade. Certainly the most important forum for this multidisciplinary field is the *International Symposium on Music Information Retrieval* (ISMIR) [11]. In 2009, ISMIR celebrates its 10<sup>th</sup> anniversary. Thus, we think it is time to look back and investigate which general topics and research problems were most important in MIR during the past decade. To this end, we analyzed the digital ISMIR proceedings [11] available online. Not only have we captured the principal topics reflected by previous, accepted ISMIR papers by means of text-based content extraction and analysis, but we have also investigated how these topics changed over time. Since MIR is a highly dynamic field of research, we

<sup>1</sup><http://www.cp.jku.at/projects/ISMIR-cloud/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

© 2009 International Society for Music Information Retrieval.

gained interesting insights, which will be detailed in the following. We further visualized the corpus of ISMIR documents via clusters of topics described by sets of terms. To this end, we employed two approaches based on *Non-Negative Matrix Factorization* (NMF) and *Principal Components Analysis* (PCA). Two prototype applications are provided as a proof-of-concept. The first is a Web application for browsing the cumulative online ISMIR proceedings theme-wise. The second is an offline OpenGL application that visualizes the *ISMIR Cloud* in three dimensions, and allows for real-time interaction such as spatial navigation and text-based search for tags.

The remainder of the paper is organized as follows. Section 2. gives an overview of related work on text information extraction and retrieval, topic-based clustering, and visualization. Section 3. describes the features we extracted from the ISMIR corpus. In Section 4., we present our approaches to visualize and browse the papers. Finally, Section 5. draws conclusions and points out directions for future work.

### 2. RELATED WORK

Related work mainly falls into the two fields of *text mining*, more precisely, text-based information extraction/retrieval and *clustering and visualizing high-dimensional data*. In line with the dedication of this paper to the MIR community, we will focus on work carried out in the context of music information research.

In the context of MIR, extracting terms from texts, more precisely, from Web documents, in order to tag a music artist has first been addressed in [28], where Whitman and Lawrence extract different term sets (e.g., noun phrases and adjectives) from artist-related Web pages. Based on term occurrences, individual term profiles are created for each artist. The authors then use the overlap between the term profiles of two artists as an estimate for their similarity. A quite similar approach is presented in [13]. Knees et al. however do not use specific term sets, but create a term list directly from the retrieved Web pages. Subsequently, a term selection technique is applied to filter out less important terms. Hereafter, the TF-IDF measure, e.g., [31], is used to weight the remaining words and subsequently

create a weighted term profile for each artist. Knees et al. propose their approach for artist-to-genre classification and similarity measurement.

A text-based music retrieval system that builds upon methods for term extraction from Web pages, term weighting, audio feature extraction, and similarity measurement is presented in [15]. In this paper, the authors relate audio features extracted from a given music collection with terms extracted from Web pages that contain part of the metadata present in the music collection. These terms are then weighted using an adapted version of the TF-IDF measure and joined with the audio features to build a feature vector for each track, which serves as a track descriptor. This approach allows for searching music collections via descriptive natural language terms, e.g., by issuing queries like “guitar riff” or “metal band with front woman”. Other work related to MIR that makes use of text mining techniques includes [10], where a POS tagger is used to search *last.fm* [17] tags for adjectives that describe the mood of a song. In [3] the machine learning algorithm *AdaBoost* is used to learn relations between acoustic features and *last.fm* tags.

As for general work on text-based information extraction and retrieval, different methods for term selection and term weighting have been analyzed with respect to their performance in text categorization, cf. [2, 16, 30], in text-based retrieval, cf. [24], and in clustering, cf. [6]. A comprehensive evaluation of term weighting techniques and similarity measures for information retrieval purposes is presented in [31]. In their extensive evaluation of various formulations of TF, IDF, and similarity measures, Zobel and Moffat conclude that no single combination outperforms the others consistently. In fact, the performance of any combination was found to be highly dependent on the domain and query set it had been applied to. Text-based IE from the Web usually relies on identifying or learning specific patterns that contain the information to be extracted. Already in [7], the use of static rules to determine hyponyms in text corpora was proposed. [4] presents a system that complements generic text patterns with domain-specific rules found by pattern extraction via search engines and subsequent selection of high-quality extraction rules. [1] proposes an approach that solely relies on Google’s page counts for specific patterns to determine instances of a given concept.

As for clustering and visualizing high-dimensional feature data in the context of MIR, in [21] *Non-Negative Matrix Factorization* (NMF) [18] was employed to determine clusters of concepts based on tags describing music artists, which were extracted from *last.fm*. Using NMF on features gained from a term weighting approach in order to cluster documents was already proposed in [29].

Another data projection and visualization technique is *Principal Components Analysis* (PCA) [8, 12]. PCA consists in a linear projection of high-dimensional data onto a small set of orthogonal dimensions with minimal loss in variance. The relative distances between data points in the high-dimensional space are preserved as good as possible in the low-dimensional projection. Reducing the dimen-

sionality of the feature space to two or three thus allows for the visualization of possible low-dimensional structure in the original high-dimensional data space.

A precedent of interactive visualization of scientific information flows (such as citation patterns across disciplines and journals, and temporal evolution of citation indices) is provided by [22, 27]. A notable difference of this approach is that the information being visualized is obtained from bibliometric data (journal citation reports) rather than data obtained through automatic content extraction from publications.

### 3. DATA ACQUISITION AND FEATURE EXTRACTION

Text-based information extraction and retrieval commonly relies on the *bag of words* model, which can be traced back at least to [19]. According to this model, a document is represented as an unordered set of its words, ignoring structure and grammar rules. Words can be generalized to terms, where a term may be a single word or a sequence of  $n$  words ( $n$ -grams), or correspond to some grammatical structure, e.g., a noun phrase. Using such a bag of words representation, each term  $t$  describing a particular document  $d$  is commonly assigned a weight  $w_{t,d}$  that estimates the importance of  $t$  in  $d$ . Each document can then be described by a *feature vector* that aggregates the single term weights. When considering a whole corpus of documents, each document can be thought of as a representation of its feature vector in a *feature space* or *vector space* whose dimensions correspond to the particular term weights. This so-called *vector space model* is a fundamental model in information retrieval and was originally described in [25]. For the term weighting function  $w_{t,d}$ , in modern information retrieval, typically some variant of TF-IDF scores is used. The TF term gives more weight to terms that appear many times in a document, whereas the IDF term ensures that less weight is given to terms that appear in many documents. More details on term weighting via TF-IDF can be found in [32]. The TF-IDF function assigns a weight  $w_{t,d}$  to a particular term  $t$  and document  $d$ . Calculating  $w_{t,d}$  for all terms remaining after having performed term selection on the terms extracted from the corpus thus yields a representation of  $d$  as a term weight vector in the feature space.

Following these basic principles of text-based information retrieval, we performed feature extraction as follows. First, we retrieved the PDF files of the accepted ISMIR submissions from the online repository [11]. This yielded effectively 719 documents. Subsequently, we converted the PDF files to standard text files. To this end, the GNU/Linux tools *pdftotext* from *xpdf-utils* and *iconv* from *libc6* were used. Minor problems encountered in the transcription process, such as occasional truncation of words, were addressed by prefix/suffix filtering as detailed later. Next, we employed the part-of-speech (POS) tagger *Geniatagger* [26] to extract all noun phrases from the corpus since we believe that these are most important to describe the content of ISMIR papers. As the output of the POS tagger

contained a lot of noise, we subsequently applied some ad-hoc filters: We discarded all terms containing non-alphabetic characters and retained only trigrams, bigrams, and unigrams. This yielded approximately 70,000 terms.

Since we aimed at emphasizing terms important to MIR, we performed term selection via Web-based filtering of terms that tend to be important in a general context. Such terms thus tend to be rather unimportant in the context of MIR and also not very discriminative for the corpus of ISMIR papers. For this purpose, we queried the Web search engine *exalead* [5] for the extracted  $n$ -grams as exact phrase and retrieved the returned page-count-values. We then discarded all terms whose TF in the ISMIR corpus was lower than their page-count-value. To alleviate the problem of truncated words after PDF-to-text-conversion, we removed any  $n$ -gram  $v$  that was a prefix/suffix of another  $n$ -gram  $w$  (of equal  $n$ ) and whose minimal TF among the single words occurring in  $v$  was lower than the minimal TF among the single words occurring in  $w$ , assuming that truncated words typically have a low TF.

This approach finally yielded a term list of approximately 12,500 terms. We extracted, for each of these terms, its absolute TF count per ISMIR document and its global DF count in the corpus. Weighting each TF value with the (logarithmic) IDF obtained from the ISMIR corpus according to Formula 1 provided a TF·IDF vector representation of each ISMIR document. In Equation 1,  $n$  is the total number of documents in the corpus,  $tf_{t,d}$  is the number of occurrences of term  $t$  in  $d$ , and  $df_t$  is the number of documents in the whole corpus in which  $t$  occurs at least once. Concatenating the TF·IDF representation of all documents yields a term-document matrix.

$$w_{t,d} = \begin{cases} tf_{t,d} \cdot \log \frac{n}{df_t} & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

#### 4. VISUALIZING ISMIR

Determining and illustrating the most important concepts tackled by ISMIR papers over time, we used Non-Negative Matrix Factorization (NMF) and Principal Components Analysis (PCA) as elaborated in the following.

##### 4.1 Finding Concepts by Non-Negative Matrix Factorization

Topic detection on the TF·IDF vectors, calculated as described in the previous section, was performed as proposed, for example, in [9, 18, 21, 29]. For NMF calculation, the cost function is the square of the Euclidean distance, and update takes place by the standard multiplicative update rules. Initialization is done randomly. NMF aims to find an approximate decomposition (into matrices  $W$  and  $H$ ) with non-negativity constraints, cf. Equation 2, where  $V$  is the  $n \times m$  matrix of the 12,500-dimensional TF·IDF vectors and  $m = 719$  documents.

$$V \approx WH \quad (2)$$

Matrix  $W$  is interpreted as containing the amount each of the  $n$  terms is associated with each of  $r$  concepts, and  $H$  as containing the amount each document is associated with each of the  $r$  concepts.

##### 4.2 Changes of ISMIR Topics over Time

The association between concepts and documents that results from NMF over the term-document matrix, allows us to make an association between concepts and years (by summing the activation of concepts in the documents of each year). Figure 1 shows the evolution of  $r = 22$  concepts over time, with the overall height representing the number of relevant publications in the year. The legend shows the three top-weighted terms for each concept. The concepts have been ordered vertically according to their growth/decline over time, the most growing concepts being on top. To this end, we performed linear regression on the development of concept weights during the considered time span.

Several interesting observations can be made. Firstly, note that the concepts seem to be of different categories. Whereas some concepts clearly represent topics (e.g., genre classification, onset detection, rhythm description, or fingerprinting), others seem to represent methods that can be used to solve different types of problems (such as matrix factorization or dynamic time-warping).

Secondly, even if the presence of most concepts is rather stable over the years, there are some notable changes over time. Some of the changes that the analysis reveals are not very surprising, such as the fact that semantic audio annotation performed via collaborative tagging (such as employed by last.fm) or Web mining, was virtually absent as a research topic in the first ISMIR conferences. By 2008, it has gained the largest share. Other changes are less obvious. For example, the share of query-by-humming/singing in ISMIR 2002 papers was considerably higher than it was in later years. Furthermore, genre classification seems to have boomed briefly at ISMIR 2005. This might be related to the MIREX 2005 genre classification contest.

##### 4.3 A Web-Interface to Concept-based search of ISMIR publications

The concepts found by NMF can also be used to create an interface for searching ISMIR papers associated with particular concepts. The simplest form of such an interface is a selection screen that lets the user select one or more of the  $r$  concepts of interest. The user selection is transformed into a vector of length  $r$ , with all entries set to zero except these that correspond to selected concepts, which are set to one. This vector is used as a query vector, and compared to each of the documents' concept vectors by cosine similarity [23]. The outcome is then presented to the user as a list of suggested documents ranked according to their similarity to the query vector. More elaborate approaches would include search refinement (e.g., by relevance feedback), or query term expansion based on the

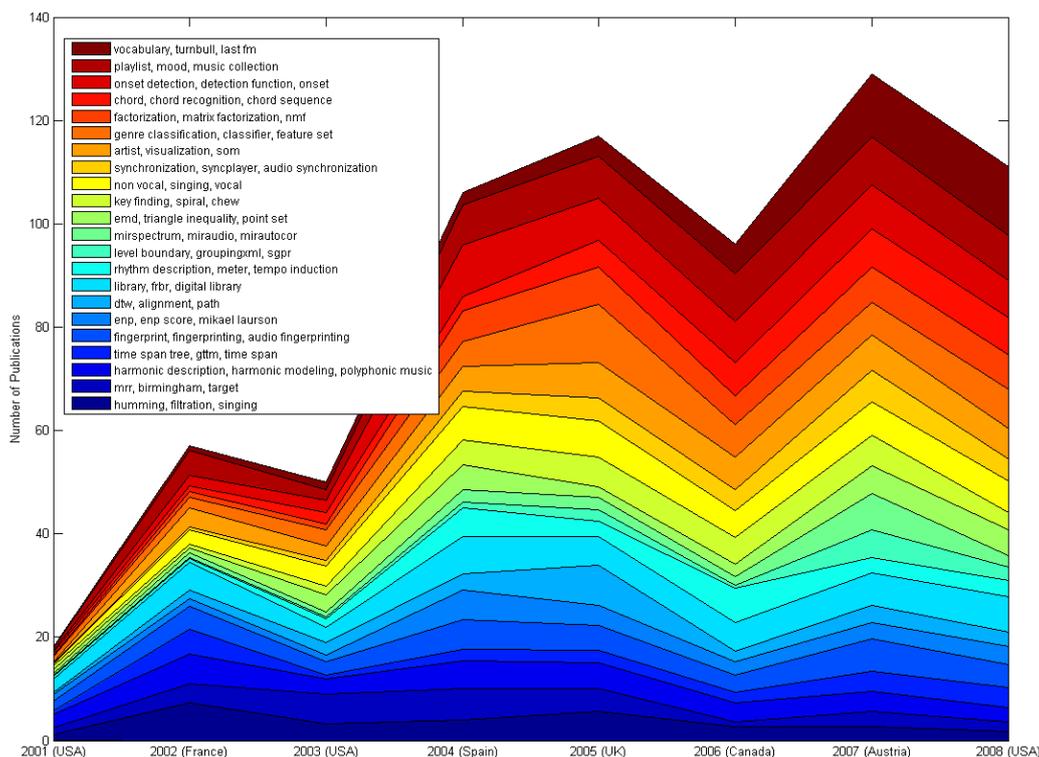


Figure 1. Evolution of the main ISMIR topics over the years.

concept vectors. We have implemented a small prototype application for document retrieval as a web service.

#### 4.4 ISMIRviewer: Navigating the ISMIR Cloud

To visualize the semantic content of the joint ISMIR publications, we pursue the idea of the *tag cloud*. In this subsection, we describe how we construct the tag cloud containing the terms extracted from the documents, and present the *ISMIRviewer*, an application for interactively navigating this tag cloud.

The term-document matrix that contains the IDF-weighted term frequency of each term in each of the 719 documents can be seen as specifying each term as a point in a 719-dimensional Euclidean space. In general, the more frequently two terms occur in the same subset of documents, the closer they will be in this space. This high-dimensional space, however, cannot be used directly for visualizing the relationships among terms. Moreover, if multiple documents contain the same terms with similar frequencies, there will be redundancies in the corresponding dimensions. In this case, the dimensionality of the feature space can be reduced without losing information about the distance between the terms and their relative location. PCA is a technique for such a dimensionality reduction, in which the data is projected on a set of orthogonal dimensions that have been rotated to maximize the variance along each dimension. The dimensions are ordered according to the data variance they hold. In this way, a subset of dimensions of any size can be chosen with maximal data variance. The principal components are obtained by computing the eigenvalues of the covariance matrix of the data, cf. [12].

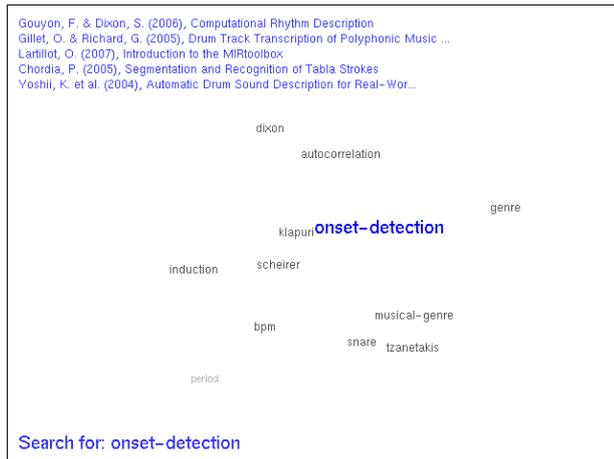
Since we aim at providing a spatial visualization, the number of dimensions is obviously limited to three. However, we found that projecting the data from 719 to three dimensions directly was not useful in this case as the first three principal components accounted only for 12% of the variance in the data (90% being reached when using 347 dimensions). When visualized, the terms are very condensed in space, where the variance is highly dominated by a few common terms like “music” and “audio”. Using the logarithms of term frequencies alleviated this problem slightly, but not satisfyingly.

Instead, we have opted for a two-stage approach to data reduction. The first stage applies NMF, as described in subsection 4.1. It yields a small set of basis vectors, which are formally activation patterns over documents, and tend to represent musically meaningful concepts. Each term  $t$  has an activation value for each concept  $c$ , denoting how relevant  $t$  is to  $c$ . Experimentation with NMF using different numbers of concepts shows that an NMF reduction to twenty concepts include most recognizable subfields of MIR without introducing many unrecognizable concepts<sup>2</sup>. Given these twenty concepts, terms are filtered to include only the 100 most activated terms for each of the concepts.

As a second stage, we perform dimensionality reduction to three dimensions through PCA on the subset of terms and the activations over the twenty concepts that were obtained in the first stage. The resulting space is less densely populated and the terms it contains tend to be more MIR-relevant.

For interactive inspection of the constructed tag cloud,

<sup>2</sup> As judged informally by the authors



**Figure 2.** Screenshot of the ISMIRviewer showing the *onset-detection* neighborhood of the ISMIR Cloud.

we have developed the *ISMIRviewer*, that allows the user to freely rotate the space and zoom in on regions using the mouse. Furthermore, subsets of the cloud can be selected by text search. As the user types, the matching tags light up. For each matching tag, neighboring tags are displayed, while remote and non-matching tags are dimmed. For the given selection of tags, five publications are shown in the corner of the screen that have been determined to be the most relevant for that term. Instead of determining document relevance through TF-IDF, the term is mapped to the documents via the concepts found by NMF. This effectively realizes a document search by *query expansion*.

In this way, the user can search for MIR-related topics, methods, or author names, and obtain relevant publications. Figure 2 shows a screenshot of the application. The displayed terms are the result of searching for the term *onset-detection*. The neighborhood of the search term (small black font) contains related concepts, e.g., *period*, *bpm*, techniques used (e.g. *autocorrelation*), and authors who have published on onset detection, such as Klapuri and Dixon.

## 5. CONCLUSIONS

In this paper, we analyzed the proceedings of the past ISMIR conferences, extracted terms from the documents, and employed text and Web mining techniques to distill a set of  $n$ -grams we believe to be important to describe the field of music information retrieval. Using a TF-IDF weighting function, we described each document by means of its term weights. We then applied clustering techniques to reveal the most important concepts covered by the ISMIR papers. Furthermore, a year-wise analysis of the publications revealed interesting changes of topics addressed in ISMIR over the years. For example, in ISMIR 2002, query-by-humming was a major topic, that has received considerably less attention in later years. Furthermore, genre classification had a particularly large share in ISMIR 2005.

Moreover, we presented two prototype applications that provide access to the semantic content of the past ISMIR

publications. The first one is a Web-based retrieval system to search the corpus of ISMIR proceedings via the concepts found by NMF. The second one, which we call *ISMIRViewer*, provides an interactive tag cloud visualization to reveal the relationships between MIR related terms. It employs a focus and context technique to show subsets of the tag cloud in response to user-entered text queries, and provides the ISMIR publications that are most relevant to the text queries.

The applications are presented as a proof-of-concept, their user-interfaces leave room for improvement. Further work to be done includes investigating other clustering techniques, e.g., *Aligned Self-Organizing Maps* [20] or *Music Description Maps* [14].

## 6. ACKNOWLEDGMENTS

This work was supported by the Austrian FWF under grants L511 and P19349.

## 7. REFERENCES

- [1] P. Cimiano, S. Handschuh, and S. Staab. Towards the Self-Annotating Web. In *Proceedings of the 13th International Conference on World Wide Web (WWW 2004)*, pages 462–471, New York, NY, USA, 2004. ACM Press.
- [2] F. Debole and F. Sebastiani. Supervised Term Weighting for Automated Text Categorization. In *Proceedings the 18th ACM Symposium on Applied Computing (SAC 2003)*, pages 784–788, Melbourne, FL, USA, March 9–12 2003. ACM.
- [3] D. Eck, T. Bertin-Mahieux, and P. Lamere. Autotagging Music Using Supervised Machine Learning. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, Vienna, Austria, September 23–27 2007.
- [4] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI 2004)*, pages 391–398, San Jose, CA, USA, 2004.
- [5] Exalead: Redefining information access for the enterprise and the web.: <http://www.exalead.com>, 2009. (access: April 2009).
- [6] V. Fresno, R. Martínez, and S. Montalvo. Improving web page clustering through selecting appropriate term weighting functions. In *Proceedings of the 1st IEEE International Conference on Digital Information Management (ICDIM 2006)*, pages 511–518, Bangalore, India, December 6–8 2006.
- [7] M. A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th*

- Conference on Computational Linguistics – Vol. 2*, pages 539–545, Nantes, France, August 1992.
- [8] H. Hotelling. Analysis of a Complex of Statistical Variables Into Principal Components. *Journal of Educational Psychology*, 24:417–441 and 498–520, 1933.
- [9] P. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.
- [10] X. Hu, M. Bay, and J. S. Downie. Creating a Simplified Music Mood Classification Ground-Truth Set. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, Vienna, Austria, September 23–27 2007.
- [11] International society for music information retrieval, ismir: Conferences, publications and related activities. <http://www.ismir.net>, 2009. (access: May 2009).
- [12] I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, NY, USA, 1986.
- [13] P. Knees, E. Pampalk, and G. Widmer. Artist Classification with Web-based Data. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR 2004)*, pages 517–524, Barcelona, Spain, October 10–14 2004.
- [14] P. Knees, T. Pohle, M. Schedl, and G. Widmer. Automatically Describing Music on a Map. In *Proceedings of 1st Workshop on Learning the Semantics of Audio Signals (LSAS 2006)*, Athens, Greece, December 6–8 2006.
- [15] P. Knees, T. Pohle, M. Schedl, and G. Widmer. A Music Search Engine Built upon Audio-based and Web-based Similarity Measures. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, Amsterdam, the Netherlands, July 23–27 2007.
- [16] M. Lan, C.-L. Tan, H.-B. Low, and S.-Y. Sung. A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web (WWW 2005)*, pages 1032–1033, Chiba, Japan, May 10–14 2005. ACM Press.
- [17] Last.fm - listen to internet radio and the largest music catalogue online. <http://last.fm>, 2008. (access: April 2009).
- [18] D. D. Lee and H. S. Seung. Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature*, 401(6755):788–791, 1999.
- [19] H. P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal*, pages 309–317, October 1957.
- [20] E. Pampalk. Aligned Self-Organizing Maps. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM 2003)*, pages 185–190, Kitakyushu, Japan, September 11–14 2003. Kyushu Institute of Technology.
- [21] T. Pohle, P. Knees, M. Schedl, and G. Widmer. Building an Interactive Next-Generation Artist Recommender Based on Automatically Derived High-Level Concepts. In *Proceedings of the 5th International Workshop on Content-Based Multimedia Indexing (CBMI'07)*, Bordeaux, France, June 25–27 2007.
- [22] M. Rosvall and C. T. Bergstrom. Maps of information flow reveal community structure in complex networks. In *Proceedings of the National Academy of Sciences USA*, volume 105, pages 1118–1123, 2007.
- [23] G. Salton. The Use of Citations as an Aid to Automatic Content Analysis. Technical Report ISR-2, Section III, Harvard Computation Laboratory, Cambridge, MA, USA, 1962.
- [24] G. Salton and C. Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513–523, 1988.
- [25] G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, 1975.
- [26] Y. Tsuruoka and J. Tsujii. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of HLT/EMNLP*, pages 467–474, 2005.
- [27] well-formed.eigenfactor.org : Visualizing information flow in science:. <http://well-formed.eigenfactor.org>, 2009. (access: May 2009).
- [28] B. Whitman and S. Lawrence. Inferring Descriptions and Similarity for Music from Community Metadata. In *Proceedings of the 2002 International Computer Music Conference (ICMC 2002)*, pages 591–598, Göteborg, Sweden, September 16–21 2002.
- [29] W. Xu, X. Liu, and Y. Gong. Document Clustering Based on Non-negative Matrix Factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, pages 267–273, Toronto, Canada, July 28–August 1 2003. ACM Press.
- [30] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, USA, 1997. Morgan Kaufman.
- [31] J. Zobel and A. Moffat. Exploring the Similarity Space. *ACM SIGIR Forum*, 32(1):18–34, 1998.
- [32] J. Zobel and A. Moffat. Inverted Files for Text Search Engines. *ACM Computing Surveys*, 38:1–56, 2006.