TEMPORAL ANALYSIS OF TEXT DATA USING LATENT VARIABLE MODELS

Lasse L. Mølgaard[*]*, Jan Larsen*

Section for Cognitive Systems
DTU Informatics
DK-2800 Kgs. Lyngby, Denmark
{llm, jl}@imm.dtu.dk

Cyril Goutte

Interactive Language Technologies
National Research Council of Canada
Gatineau, Canada, QC J8X 3X7
Cyril.Goutte@cnrc-nrc.gc.ca

ABSTRACT

Detecting and tracking of temporal data is an important task in multiple applications. In this paper we study temporal text mining methods for Music Information Retrieval. We compare two ways of detecting the temporal latent semantics of a corpus extracted from Wikipedia, using a stepwise Probabilistic Latent Semantic Analysis (PLSA) approach and a global multiway PLSA method. The analysis indicates that the global analysis method is able to identify relevant trends which are difficult to get using a step-by-step approach. Furthermore we show that inspection of PLSA models with different number of factors may reveal the stability of temporal clusters making it possible to choose the relevant number of factors.

1. INTRODUCTION

Music Information Retrieval (MIR) is a multifaceted field, which until recently mostly focused on audio analysis. The use of textual descriptions, beyond using genres, has grown in popularity with the advent of different music websites, e.g. "Myspace.com", where abundant data about music has become easily available. This has for instance been investigated in [1], where textual descriptions of music were retrieved from the Web to find similarity of artists. The unstructured data retrieved using web crawling produces a lot of data, which requires cleaning to produce terms that actually describe musical artists and concepts. Community-based music web services such as tagging based systems, e.g. Last.fm, have also shown to be a good basis for extracting latent semantics of musical track descriptions [2]. Despite these initial efforts it is still an open question how textual data is best used in MIR. The text-based methods have so far only considered text data without any structured knowledge. In this study we investigate if the incorporation of time information in latent factor models enhances the detection and description of topics.

Tensor methods in the context of text mining have recently received some attention using higher-order decomposition methods such as the PARAllel FACtors model [3] that can be seen as a generalization of Singular Value Decomposition in higher dimensional arrays. The article [3] applies tensor decomposition methods successfully for topic detection in e-mail correspondence over a 12 month period. The article also employs a non-negatively constrained PARAFAC model forming a Nonnegative Tensor Factorization analogous to the well-known Nonnegative Matrix Factorization (NMF) [4].

Probabilistic Latent Semantic Analysis (PLSA) [5] and NMF have successfully been applied in many text analysis tasks to find interpretable latent factors. The two methods have been shown to be equivalent [6], where PLSA has the advantage of providing a direct probabilistic interpretation of the latent factors. Our work therefore investigates the extension of PLSA to tensors.

2. TEMPORAL TOPIC DETECTION

Detecting latent factors or topics in text using NMF and PLSA has assumed an unstructured and static collection of documents.

Extracting topics from a temporally changing text collection has received some attention lately, for instance by [7] and also touched by [8]. These works investigate text streams that contain documents that can be assigned a timestamp y. The timestamp may for instance be the time a news story was released, or in the case of articles describing artists it can be a timespan indicating the active years of the artist. Finding the evolution of topics over time requires assigning documents $d_1, d_2, ..., d_m$ in the collection to time intervals $y_1, y_2, ..., y_l$, as illustrated in figure 1. In contrast to the temporal topic detection approach in [7], we can assign documents to multiple time intervals, e.g. if the

[*]First author performed the work while visiting NRC Interactive Language Technologies group.

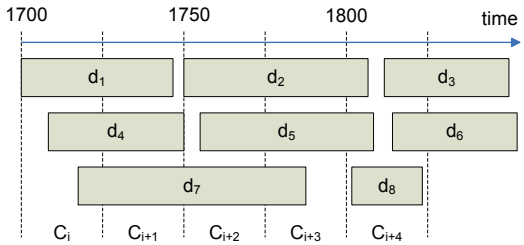

Fig. 1. An example of assigning a collection of documents d_i based on the time intervals the documents belong to. The assignment produces a document collection C_k for each time interval

active years of an artist spans more than one of the chosen time intervals. The assignment of documents then provides l sub-collections $C_1, C_2, ..., C_l$ of documents.

The next step is to extract topics and track their evolution over time.

2.1. Stepwise temporal PLSA

The approaches to temporal topic detection presented in [7] and [8] employ latent factor methods to extract distinct topics for each time interval, and then compare the found topics at succeeding time intervals to link the topics over time to form temporal topics.

We extract topics from each sub-collection C_k using a PLSA-model [5]. The model assumes that documents are represented as a bags-of-words where each document d_i is represented by an n-dimensional vector of counts of the terms in the vocabulary, forming an $n \times m$ term by document matrix for each sub-collection C_k. PLSA is defined as a latent topic model, where documents and terms are assumed independent conditionally over topics z:

$$P(t,d)_k = \sum_z^Z P(t|z)_k P(d|z)_k P(z)_k \tag{1}$$

This model can be estimated using the Expectation Maximization (EM) algorithm, cf. [5].

The topic model found for each document sub-collection C_k with parameters, $\theta_k = \{P(t|z)_k, P(d|z)_k, P(z)_k\}$, need to be stringed together with the model for the next time span θ_{k+1}. The comparison of topics is done by comparing the term profiles $P(t|z)_k$ for the topics found in the PLSA model. The similarity of two profiles is naturally measured using the KL-divergence,

$$D(\theta_{k+1}||\theta_k) = \sum_t p(t|z)_{k+1} \log \frac{p(t|z)_{k+1}}{p(t|z)_k}. \tag{2}$$

Determining whether a topic is continued in the next time span is quite simply chosen based on a threshold λ, such that two topics are linked if $D(\theta_{k+1}||\theta_k)$ is smaller than a fixed threshold λ. In this case the asymmetric KL-divergence is used in accordance with [7]. The choice of the threshold must be tuned to find the temporal links that are relevant.

2.2. Multiway PLSA

The method presented above is useful to some extent, but does not fully utilize the time information that is contained in the data. Some approaches have used the temporal aspect more directly, e.g. [9] where an incrementally trainable NMF-model is used to detect topics. This approach does include some of the temporal knowledge but still lacks global view of the important topics viewed over the whole corpus of texts.

Using multiway models, also called tensor methods we can model the topics directly over time. The 2-way PLSA model in 1 can be extended to a 3-way model by also conditioning the topics over years y, as follows:

$$P(t,d,y) = \sum_z P(t|z)P(d|z)P(y|z)P(z) \tag{3}$$

The model parameters are estimated using maximum likelihood using the EM-algorithm, e.g. as in [10]. The expectation step evaluates $P(z|t,d,y)$ using the estimated parameters at step t.

$$\text{(E-step):} \qquad P(z|t,d,y) = \frac{p(t|z)p(d|z)p(y|z)p(z)}{\sum_{z'} p(t|z')p(d|z')p(y|z')p(z')} \tag{4}$$

The M-step then updates the parameter estimates.

$$\text{(M-step):} \qquad P(z) = \frac{1}{N} \sum_{tdy} x_{tdy} P(z|t,d,y) \tag{5}$$

$$P(t|z) = \frac{\sum_{dy} x_{tdy} P(z|t,d,y)}{\sum_{tdy} x_{tdy} P(z|t,d,y)} \tag{6}$$

$$P(d|z) = \frac{\sum_{ty} x_{tdy} P(z|t,d,y)}{\sum_{tdy} x_{tdy} P(z|t,d,y)} \tag{7}$$

$$P(y|z) = \frac{\sum_{td} x_{tdy} P(z|t,d,y)}{\sum_{tdy} x_{tdy} P(z|t,d,y)} \tag{8}$$

The EM algorithm is guaranteed to converge to a local maximum of the likelihood. The EM algorithm is sensitive to initial conditions, so a number of methods to stabilize the estimation have been devised, e.g. Deterministic Annealing [5]. We have not employed these but instead rely on restarting the training procedure a number of times to find a good solution.

The time complexity of the two PLSA approaches of course depends on the number of iterations for the method to converge. Basically the most expensive operation is the

E-step of the algorithms. The cost of each iteration for 2-way PLSA is $\mathcal{O}(RZ)$ which is calculated for each of the K time steps. R is the number of non-zeros in the term-doc matrix. Each iteration for multiway PLSA (mwPLSA), takes $\mathcal{O}(RZK)$ as all time-steps are calculated simultaneously. In our experiments the algorithms typically converge in 40-50 iterations. However, the 2-way PLSA does have the advantage that the individual time steps can be calculated in parallel giving a speed-up proportional to K.

2.3. Topic model interpretation

The latent factors z of the model can be seen as topics that are present in the data. The parameters of each topic can be used as descriptions of the topic. $P(t|z)$ represents the probabilities of the terms for the topic z, thus providing a way to find words that are representative of the topic. The most straightforward method to find these keywords is to use the words with the highest probability $P(t|z)$. This approach unfortunately is somewhat flawed as the histogram reflects the overall frequency of words, which means that generally common words tend to dominate the $P(t|z)$.

This effect can be neutralized by measuring the relevance of words in a topic relative to the probability in the other topics. Measuring the difference between the histograms for each topic can be measured by use of the symmetrized Kullback-Leibler divergence:

$$KL(z, \neg z) = \sum_t \underbrace{(P(t|z) - P(t|\neg z)) \log \frac{P(t|z)}{P(t|\neg z)}}_{w_t} \quad (9)$$

This quantity is a sum of contributions from each term t, w_t. The terms that contribute with a large value of w_t are those that are relatively more special for the topic z. w_t can thus be used to choose the keywords. The keywords should be chosen from the terms that have a positive value of $P(t|z) - P(t|\neg z)$ and with the largest w_t.

3. WIKIPEDIA DATA

In this experiment we investigated the description of composers in Wikipedia. This should provide us with a dataset that spans a number of years, and provides a wide range of topics. We performed the analysis on the Wikipedia data dump saved 27th of July 2008, retrieving all documents that Wikipedians assigned to composer categories such as "Baroque composers" and "American composers". This produced a collection of 7358 documents, that were parsed so that only the running text was kept.

Initial investigations in music information web mining showed that artist names can heavily bias the results. Therefore words occurring in titles of documents, such as *Wolfgang Amadeus Mozart*, are removed from the text corpus, i.e. occurrences of the terms 'wolfgang', 'amadeus', and 'mozart' were removed from all documents in the corpus. Furthermore we removed irrelevant stopwords based on a list of 551 words. Finally terms that occurred fewer than 3 times counted over the whole dataset and terms not occurring in at least 3 different documents were removed.

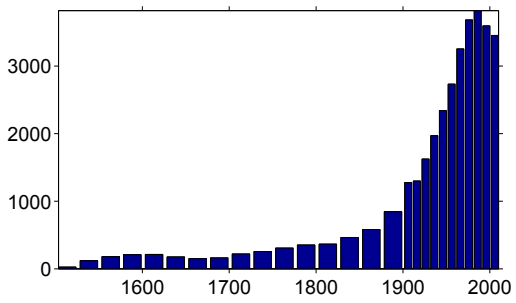

Fig. 2. Number of composer documents assigned to each of the chosen time spans.

The document collection was then represented using a bag-of-words representation forming a term-document matrix $\mathbf{X}$ where each element x_{td} represents the count of term t in document d. The vector $\mathbf{x}_d$ thus represents the term histogram for document d.

To place the documents temporally the documents were parsed to find the birth and death dates. These data are supplied in Wikipedia as documents are assigned to categories such as "1928 births" and "2007 deaths". The dataset contains active composers from around 1500 until today. The next step was then to choose the time spans to use. Inspection of the data revealed that the number of composers before 1900 is quite limited so the artists were assigned to time intervals of 25 years, giving a first time interval of [1501-1525]. After 1900 the time intervals were set to 10 years, for instance [1901-1910]. Composers were assigned to time intervals if they were alive in some of the years. We estimated the years composers were active by removing the first 20 years of their lifetime. The resulting distribution of documents on the resulting 27 time intervals is seen in figure 2.

The term by document matrix was extended with the time information by assigning the term-vector for each composer document to each era, thus forming a 3-way tensor containing terms $\times$ documents $\times$ years. The tensor was further normalized over years, such that the weight of the document summed over years is the same as in the initial term doc-matrix. I.e. $P(d) = \sum_{t,y} X_{tdy} = \sum_t X_{td}$. This was done to avoid long-lived composers dominating the resulting topics.

The resulting tensor $\mathbf{X} \in \mathbb{R}^{m \times n \times l}$ contains 18536 terms x 7358 documents x 27 time slots with 4,038,752 non-zero entries (0.11% non-zero entries).

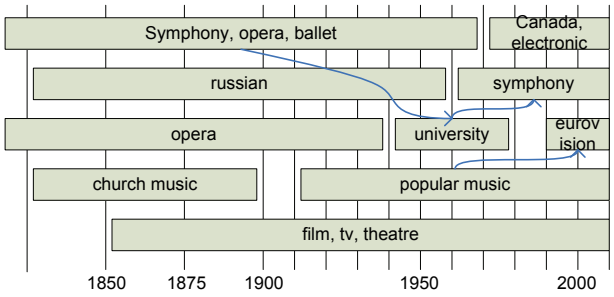

Fig. 3. Topics detected using step-by-step PLSA. The topics are depicted as connected boxes, but are the results of the KL-divergence-based linking between time slots

3.1. Term weighting

The performance of machine learning approaches in text mining often depends heavily on the preprocessing steps that are taken. Term weighting for LSA-like methods and NMF have thus shown to be paramount in getting interpretable results. We applied the well-known $tfidf$ weighting scheme, using $tf = \log(1 + x_{tdy})$ and the log-entropy document weighting, $idf = 1 + \sum_{d=1}^{D} \frac{h_{td} \log h_{td}}{log D}$, where $h_{td} = \frac{\sum_y x_{tdy}}{\sum_{dy} x_{tdy}}$. The log local weighting minimizes the effect of very frequent words, while the entropy global weight tries to discriminate important terms from common ones. The documents in Wikipedia differ quite a lot in length, therefore we employ document normalization to avoid that long articles dominate the modeled topics.

4. EXPERIMENTS

We performed experiments on the Wikipedia composer data using the stepwise temporal PLSA method and the multiway-PLSA methods.

4.1. Stepwise temporal PLSA

The step-by-step method was trained with 5 and 16 topic PLSA models for each of the l sub-collections of documents described above. The PLSA models for each time span was trained using a stopping criterion of 10^{-5} relative change of the cost function, restarting the training 10 times for each model, choosing the model minimizing the likelihood. The two setups were chosen to have one model that extracts general topics and a model with a higher number of components that detects more specific topics. The temporal topics are produced by coupling the topics at time k and $k + 1$ if the KL-divergence between the topic term distributions, $D(\theta_{k+1}|\theta_k)$, is below a threshold λ. This choice of threshold produces a number of topics that stretch over several time spans. A low setting for λ may leave out important relations, while a higher setting produces too many links to be interpretable. Figure 3 shows the topics found for the 20th century using the 5 component models. There are clearly 4 topics that are present throughout the whole period. The topics are film and TV music composers, which in the beginning contains Broadway/theater composers. The other dominant topic describes hit music composers. Quite interestingly this topic forks off a topic describing Eurovision song contest composers in the last decades.

Even though the descriptions of artists in Wikipedia contain a lot of bibliographical information it seems that the latent topics have musically meaningful keywords. As there was no use of a special vocabulary in the preprocessing phase, it is not obvious that these musically relevant phrases would be found.

The stepwise temporal PLSA approach has two basic shortcomings. The first is the difficulties in adjusting the threshold λ to find meaningful topics over time. The second is how to choose the number of components to use in the PLSA in each time span. The 5 topics that are used above do give some interpretable latent topics in the last decade as shown in figure 3. On the other hand the results for the earlier time spans (that contain less data), means that the PLSA model finds some quite specific topics at these time spans. As an example the period 1626-1650 has the following topics:

1626-1650 41%	34%	15%	8.7%	1.4%
keyboard	madrigal	viol	baroque	anglican
organ	baroque	consort	italy	liturgi
surviv	motet	lute	poppea	prayer
italy	continuo	england	italian	respons
church	monodi	charles	lincoronazion	durham
nuremberg	renaissance	royalist	opera	english
choral	venetian	masqu	finta	chiefli
baroque	style	fretwork	era	england
germani	cappella	charles's	venice	church
collect	itali	court	teatro	choral

The topics found here are quite meaningful in describing the baroque period, as the first topic describes the church music, and the second seems to find the musical styles, such as madrigals and motets. The last topic on the other hand only has a topic weight of $P(z) = 1.4\%$. This tendency was even more distinct when using 16 components in each time span.

4.2. Multi-way PLSA

We then model the full tensor, described in section 3, using the mwPLSA model. Analogously with the stepwise temporal PLSA model we stopped training reaching a change of less than 10^{-5} of the cost function. The main advantage of the mwPLSA method is that the temporal linking of topics is accomplished directly through the model estimation. The time evolution of topics can be visualized using the parameter $P(y|z)$ that gives the weight of a component for each time span. Figure 4 shows the result for 4, 8, 16,

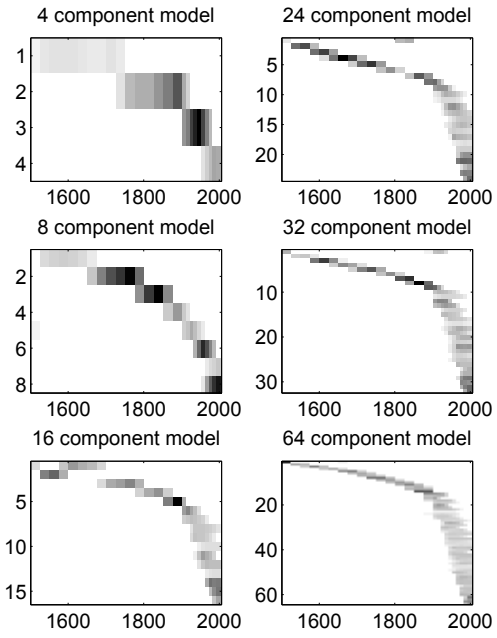

Fig. 4. Time view of components extracted using mwPLSA, showing the time profiles $P(y|z)$ as a heatmap. A dark color corresponds to higher values of $P(y|z)$

32 and 64 components as a "heatmap", where darker colors correspond to higher values of $P(y|z)$. The topics are generally unimodally distributed over time so the model only finds topics that increase in relevance and then dies off without reappearing. The skewed distribution of documents over time which we described earlier emerges clearly in the distribution of topics, as most of the topics are placed in the last century. Adding more topics to the model has two effects considering the plots. Firstly the topics seem to be sparser in time, finding topics that span fewer years. Using more topics decomposes the last century into more topics that must be semantically different as they almost all span the same years. Below we inspect the different topics to show how meaningful they are. The keywords extracted using the method mentioned above are shown in table 1, showing 5 of the topics extracted by the 32 component model, including the time spans that they belong to.

The first topic shown in table 1 is one of the two topics that accounts for the years 1626-1650, the keywords summarize the five topics found using mwPLSA. The second topic has the keywords *ragtime* and *rag*, placed in the years 1876-1921, which aligns remarkably well with the genre description on Wikipedia: *"Ragtime [...] is an originally American musical genre which enjoyed its peak popularity between 1897 and 1918."*[1]. The stepwise PLSA approach did also have *ragtime* as keywords in the 16 component

[1]http://en.wikipedia.org/wiki/Ragtime

1601-1700 2.10%	1676-1776 2.40%	1876-1921 4.70%	1921-1981 4.80%	1971-2008 6.40%
baroque	baroque	ragtime	concerto	single
italian	opera	sheet	nazi	chart
church	sonata	rag	war	album
continuo	italian	weltemignon	symphony	release
survive	court	nunc	piano	hit
court	harpsichord	ysa	ballet	track
organ	italy	schottisch	neoclassical	sold
motet	violinist	dimitti	neoclassic	demo
madrigal	church	blanch	choir	fan
cathedral	organ	parri	hochschul	pop

Table 1. Keywords for 5 of 32 components in a mwPLSA model. The assignment of years is given from $P(y|z)$ and percentages placed at each column are the corresponding component weights, $P(z)$

model, appearing as a topic from 1901-1920. The next topic seems to describe World War II, but also contains the neoclassical movement in classical music. The 16 component stepwise temporal PLSA approach finds a number of topics from 1921-1940 that describe the war, such as a topic in 1921-1930 with keywords: *war, time, year, life, influence* and two topics in 1931-1941, *1. time, war, year, life, style* and *2: theresienstadt, camp, auschwitz, deport, concentration, nazi*. These are quite unrelated to music, so it is evident that the global view of topics employed in the mwPLSA-model identifies neoclassicism to be the important keywords compared to topics from other time spans.

Some of the topics do overlap in time, such as the first two presented in table 1, and it is clear that they present different aspects of the music in the Baroque era, one representing church music (organ and madrigals), while the other describes opera and sonatas. So the overlapping topics can show how genres evolve.

5. MULTI-RESOLUTION TOPICS

The use of different number of components in the mwPLSA model, as seen in figure 4, shows that the addition of topics to the model shrinks the number of years they span. The higher specificity of the topics when using more components gives a possibility to "zoom" in on interesting topic, while the low complexity models can provide the long lines in the data.

To illustrate how the clusters are related as we add topics to the model, we can generate a so-called clusterbush, as proposed in [11]. The result for the mwPLSA-based clustering is shown in figure 5. The clusters are sorted such that the clusters placed earliest in time are placed left. It is evident the clusters related to composers from the earlier centuries form small clusters that are very stable, while the later components are somewhat more ambiguous. The clusterbush could therefore be good tool for exploring the topics at different timespans to get an estimate of the number of

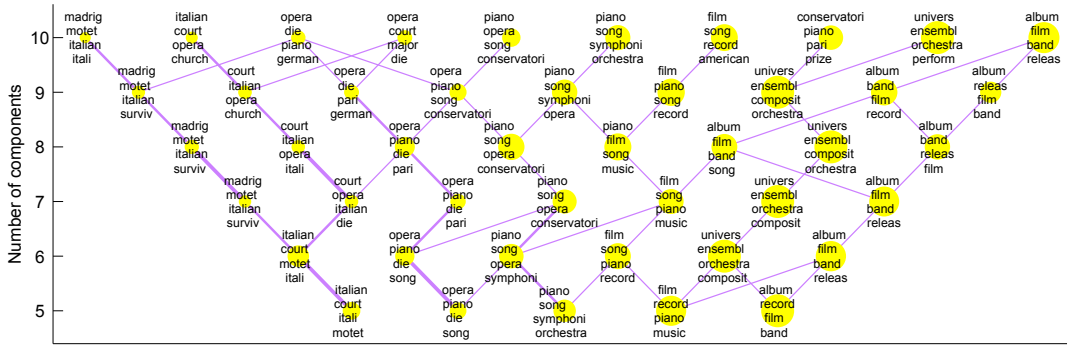

Fig. 5. "Cluster bush" visualisation of the results of the mwPLSA clustering of composers. The size of the circles correspond to the weight of the cluster, and the thickness of the line between circles how related the clusters are. Each cluster is represented by the keywords and is placed according to time from left to right.

components needed to describe the data.

6. CONCLUSION

We have investigated the use of time information in analysis of musical text in Wikipedia. It was shown that the use of time information produces meaningful latent topics, which are not readily extractable from this text collection without any prior knowledge. The stepwise temporal PLSA approach is quite fast to train and processing for each time span can readily be processed in parallel. The model, however, has the practical drawback that it requires a manual tuning of the linking threshold, and the lack of a global view of time in the training misses some topics. The multiway PLSA was shown to provide a more flexible and compact representation of the temporal data than stepwise temporal PLSA method. The flexible representation of the multiway model is a definite advantage, for instance when data has a skewed distribution as in the work resented here. The global model would also make it possible to do model selection over all time steps directly. The use of Wikipedia data also seems to be a very useful resource for semi-structured data for Music Information Retrieval that could be investigated further to harness the full potential of the data.

ACKNOWLEDGEMENTS

This work is supported by the Danish Technical Research Council, through the framework project 'Intelligent Sound', www.intelligentsound.org (STVF No. 26-04-0092).

7. REFERENCES

[1] P. Knees, E. Pampalk, and G. Widmer, "Artist classification with web-based data," in *Proceedings of ISMIR*, Barcelona, Spain, 2004.

[2] M. Levy and M. Sandler, "Learning latent semantic models for music from social tags," *Journal of New Music Research*, vol. 37, no. 2, pp. 137–150, 2008.

[3] B. Bader, M. Berry, and M. Browne, *Survey of Text Mining II Clustering, Classification, and Retrieval*, chapter Discussion tracking in Enron email using PARAFAC, pp. 147–163, Springer, 2008.

[4] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[5] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proc. 22nd Annual ACM Conf. on Research and Development in Information Retrieval*, Berkeley, California, August 1999, pp. 50–57.

[6] E. Gaussier and C. Goutte, "Relation between plsa and nmf and implications," in *Proc. 28th annual ACM SIGIR conference*, New York, NY, USA, 2005, pp. 601–602, ACM.

[7] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," in *Proc. of KDD 05*. 2005, pp. 198–207, ACM Press.

[8] M. W. Berry and M. Brown, "Email surveillance using non-negative matrix factorization," *Computational and Mathematical Organization Theory*, vol. 11, pp. 249–264, 2005.

[9] B. Cao, D. Shen, J-T Sun, X. Wang, Q. Yang, and Z. Chen, "Detect and track latent factors with online nonnegative matrix factorization," in *Proc. of IJCAI-07*, Hyderabad, India, 2007, pp. 2689–2694.

[10] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 435–447, Feb. 2008.

[11] F. Å. Nielsen, D. Balslev, and L. K. Hansen, "Mining the posterior cingulate: Segregation between memory and pain components," *NeuroImage*, vol. 27, no. 3, pp. 520–532, 2005.