VISUALLY SUMMARIZING THE EVOLUTION OF DOCUMENTS UNDER A SOCIAL TAG

André Gohr
Leibniz Institute of Plant Biochemistry, Weinbergweg 3, 06120 Halle, Germany

Myra Spiliopoulou
Institute of Computer Science, Otto-von-Guericke University, 39016 Magdeburg, Germany

Alexander Hinneburg
Institute of Computer Science, Martin-Luther University Halle-Wittenberg, 06099 Halle, Germany

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Keywords: Knowledge management, Visualization, Social web, Tag semantics, Topic modeling.

Abstract: Tags are intensively used in social platforms to annotate resources: Tagging is a social phenomenon, because users do not only annotate to organize their resources but also to associate semantics to resources contributed by third parties. This leads often to semantic ambiguities: Popular tags are associated with very disparate meanings, even to the extend that some tags (e.g. ”beautiful” or ”toread”) are irrelevant to the semantics of the resources they annotate. We propose a method that learns a topic model for documents under a tag and visualizes the different meanings associated with the tag.
Our approach deals with the following problems. First, tag miscellany is a temporal phenomenon: tags acquire multiple semantics gradually, as users apply them to disparate documents. Hence, our method must capture and visualize the evolution of the topics in a stream of documents. Second, the meanings associated to a tag must be presented in a human-understandable way; This concerns both the choice of words and the visualization of all meanings. Our method uses AdaptivePLSA, a variation of Probabilistic Latent Semantic Analysis for streams, to learn and adapt topics on a stream of documents annotated with a specific tag. We propose a visualization technique called Topic Table to visualize document prototypes derived from topics and their evolution over time. We show by a case study how our method captures the evolution of tags selected as frequent and ambiguous, and visualizes their semantics in a comprehensible way. Additionally, we show the effectiveness by adding alien resources under a tag. Our approach indeed visualizes hints to the added documents.

1 INTRODUCTION

Collaborative tagging systems provide a popular way to share and organize various resources including bibliographic entries describing academic articles. Examples are Bibsonomy[1], CiteYouLike[2] and Connotea[3]. Quoting Golder and Huberman (Golder and Huberman, 2006, pp. 200 and 203) ”tagging is fundamentally about sensemaking”, it is ”an act of organizing through labeling, a way of making sense of many discrete, varied items [resources] according to their meaning.” Aiming at organizing resources, collaborative tagging systems assist users at two levels. First, at user level, they help users to organize their own documents. Second, at community level, such systems allow users to find interesting resources contributed by other users through searching by tags. To work effectively at community level, two assumptions must hold: (i) users assign tags in a coherent way and (ii) users are capable to deduce the meanings of resources from tags. In real life, often both assumptions are violated because collaborative tagging systems are not centrally managed.

[1]http://www.bibsonomy.org
[2]http://www.citeulike.org
[3]http://www.connotea.org

Gohr A., Spiliopoulou M. and Hinneburg A..
VISUALLY SUMMARIZING THE EVOLUTION OF DOCUMENTS UNDER A SOCIAL TAG.
DOI: 10.5220/0003096100850094
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 85-94
ISBN: 978-989-8425-28-7
Copyright © 2010 SCITEPRESS (Science and Technology Publications, Lda.)

We identify two situations that might be unsatisfying for particular users: (i) tags with multiple semantics and (ii) unfamiliar tags without clear meaning. Widely used tags might have multiple semantics (Suchanek et al., 2008), some of which a particular user might not be aware of. That ambiguity of tags is likely to be promoted by the very fact that tagging is a social activity: if a "leader" user (Goyal et al., 2008) favors a specific tag for a document, other users may decide to use the same tag for documents even when those documents diverge in content. Second, tags that are unfamiliar to a user cannot be effectively used for organizing and searching resources. If a user wants to learn about the meanings of an ambiguous or unknown tag, the user might inspect a sample of the resources annotated with that tag. But inspecting whole documents is time consuming especially when the tag is frequently used.

Therefore, we propose a new method, comprised of an unsupervised learner and a visualization technique. The unsupervised learner is based on probabilistic topic modeling and discovers topics associated with a given tag on the foundation of documents annotated with it. From the topics we derive document prototypes that are presented by the proposed visualization technique. Thereby, the visualization summarizes the documents' contents at a glance allowing a user to get an overview about the meanings of the tag itself.

The appearance of ambiguous meanings of a tag can be also a temporal process, e.g. methods developed for social network analysis are later applied and modified to analyze gene regulatory networks and documents from both research areas are annotated with the tag *network*. Consequently, our method summarizes the evolution of document prototypes under a tag through time.

We integrate two kinds of information into the visualization, namely the document prototypes as well as their evolution and relative strength over time. The challenges of the design of a new visualization technique are (i) to use the canvas efficiently and (ii) display the dominant information (main document prototypes) while retaining less dominant information.

Document prototypes are derived from topics learned by probabilistic latent semantic analysis (PLSA) (Hofmann, 2001). We capture the evolution of topics by AdaptivePLSA (Gohr et al., 2009), an extension of PLSA for dynamic topic modeling. AdaptivePLSA learns a series of PLSA topic models over time and effectively prevents label switching meaning the k^{th} topic at the $(i+1)^{\text{th}}$ time point evolves from the topic k at the previous time point i. This makes AdaptivePLSA especially useful to extract topics over time in an intuitive manner.

We introduce document prototypes for document collection summarization, explain how collaborative tagging systems capture user interactions and describe how we construct a stream of documents under a tag in Section 3. These streams are used to learn topics over time as we briefly review in Section 4. In Section 5, we introduce our new visualization technique and explain its features in detail. In Section 6, we present a case study using data from the collaborative tagging system Bibsonomy and show the effectiveness of our approach.

2 RELATED WORK

In topic modeling literature (Hofmann, 2001; Blei et al., 2003), topics, which are not predefined but learned from documents and represented as discrete probability distributions over the vocabulary, are often presented by listing most likely words. Additional pieces of information like the relative strength of topics are neglected. Research (Boyd-Graber et al., 2009; Mei et al., 2007) to enhance presentation of topics for human inspection suggests to present words that not necessarily have to be the most likely words but the most descriptive words for a topic. To report topics learned by dynamic topic modeling, (Blei and Lafferty, 2006) list the most likely words for topics at several time points. Additionally, they plot the probability of certain words for a topic at different time points to give hints about how this topic changes through time. We propose Topic Table that deals with any type of word lists for topics and visualizes all topics and their relative strength through time at a glance.

ThemeRiver (Havre et al., 2002) uses a river metaphor to visualize changes in document contents over time but it relies on manually predefined words for which it visualizes their document frequencies[4] at several time points. Curved flows to whose widths the frequencies are mapped visualize their change through time. Space of the canvas is wasted whenever the width of the ThemeRiver is small and integration of text into narrow curved flows is difficult. Applying ThemeRiver to report topics over time with additional pieces of information as Topic Table does is not straight forward because ThemeRiver aims at presenting other kinds of information changing over time.

We apply Topic Table and AdaptivePLSA to summarize document contents of a social tagging system under tags. Tags reflect user interests and can

[4]Number of documents containing the word.

be exploited to assist users A common assumption is that tags are representative of resource semantics. Recently, this assumption started being questioned. Quoting Zanardi and Capra "... as tags are informally defined, continually changing, and ungoverned , social tagging has often been criticized for lowering, rather than increasing, the efficiency of searching ..." (Zanardi and Capra, 2008, p. 51). Hence, it seems reasonable to use tags, especially popular ones, with caution. Nonetheless, we believe that visually summarizing contents of documents that people associate with such tags helps users to learn the multiple meanings of a tag or to understand for what resources a yet unknown tag is used. Summarization and visualization of document contents is a way of knowledge generation in collaborative tagging systems which subsumes perspectives on a tag of many different users. The study of content to assess tag semantics is not by itself new. For example, (Moxley et al., 2009) derive semantics of tags assigned to Flickr pictures by analyzing geographical coordinates of the depicted locations. However, we also account for the fact that the meaning(s) associated with a tag may change over time.

Beside AdaptivePLSA (Gohr et al., 2009) that extends PLSA (Hofmann, 2001) to streaming document collections and that is used in this study, other approaches (Mei and Zhai, 2005; Blei and Lafferty, 2006; Wang and McCallum, 2006) model dynamic document collections, too. Some allow for words to become obsolete and irrelevant while others emerge (AlSumait et al., 2008; Chou and Chen, 2008). Capturing terminological evolution is indispensable for visualizing the semantic evolution of tags, because that evolution is inevitably associated with the increased importance of some words that were irrelevant or unknown in the past.

3 SUMMARIZING DOCUMENTS

The aim is to provide users of collaborative tagging systems with a summary of contents under tags by document prototypes so that these users, if in doubt about the meaning and usage of a certain tag, might inspect this summary to clarify its meaning.

3.1 Document Prototypes

The contents of a document collection $\vec{D}$ can be summarized by prototypes of documents in $\vec{D}$. Document prototypes abstract from the documents and thereby describe the whole set of documents in a condensed way. Thus, inspecting them allows to get an overview about the contents of the documents. Because their number is much smaller than the number of documents, inspecting prototypes is more efficient than reading single documents. We denote the collection of documents by the vector $\vec{D}$ of document IDs to allow for multiple occurrences of documents.

We use probabilistic topic modeling of documents to derive document prototypes. Topic modeling often assumes topics to be represented by multinomial distributions over words of the vocabulary (Hofmann, 2001; Blei et al., 2003). Topics capture patterns of words that often co-occur in different documents.

Because topics are distributions over words they are less suitable to summarize document collections. But being a word distribution a topic allows to rank words according to their probability. The top ranked words are most strongly associated to that topic. Inspecting these words allows to deduce what the topic's meaning. Consequently, we define for each learned topic a document prototype consisting of the N_{top} top ranked words for that topic.

Many collections change over time, because, for example, new documents are added. As an example of such a collection, consider the documents associated with a certain tag in a collaborative tagging system. As users interact with such systems, they contribute new documents over time and tag these documents. To provide a summary of contents of documents annotated with a tag, we determine document prototypes over time. In contrast to summarizing static collection, we would also have to derive how these prototypes change over time to capture the dynamic nature of these collections.

We adapt the approach of summarizing a static document collection by examining the collection as it evolves over time. Therefore, we define a stream of documents and learn topics for successive parts of the stream using an extension of probabilistic latent semantic analysis (Hofmann, 2001) described in Section 4. From the topics over time we derive document prototypes over time to be visualized for studying how the contents of documents change through time. But first, we elaborate on how collaborative tagging systems are used for managing annotations of documents with tags. Next, we explain how we construct a stream of documents under a tag to study how that content changes over time.

3.2 Tagging Events in Collaborative Tagging Systems

Collaborative tagging systems for academic articles manage bibliographic entries that are contributed by users. Bibliographic entries contain author informa-

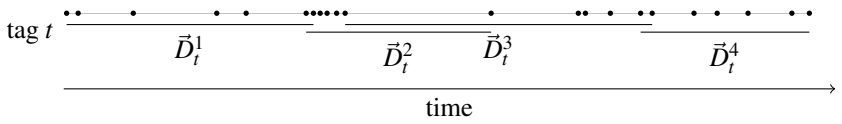

Figure 1: Stream of tagging events (small black dots) according to tag t. Each tagging event assigns tag t to one document. Positions of the sliding window of length $l = 7$ are represented by horizontal lines. The sliding window shifts forward by $l_{new} = 5$ new tagging events (tagged documents) and it covers at its four positions the following document subsequences $\vec{D}_t^1, \vec{D}_t^2, \vec{D}_t^3$ and $\vec{D}_t^4$.

tion, the title and the abstract. We use the abstract substitutional for the content of the corresponding full article because articles are often not available due to copyrights. In the sequel, we term these abstracts documents. In addition, collaborative tagging systems manage tagging events. Tags are short descriptors defined by users and can be arbitrarily assigned to documents. A tagging event is an annotation of a bibliographic entry – and hence of the corresponding document – with a certain tag by a particular user at some time point.

In this study, we neglect the information about which particular user has assigned a tag to a document. Thus, a tagging event is a triple (t, d, τ) of a tag t, a document with ID d and a time stamp τ.

3.3 Document Stream under a Tag

Time stamps of tagging events induce an ordering on documents $\vec{D}_t = \langle d_1, \ldots, d_{N_t} \rangle$ annotated with a tag t. We call all documents of $\vec{D}_t$ documents under tag t. The stream $\vec{D}_t$ may include identical documents if these have been annotated multiple times by different users with tag t. Document contents of the sequence $\vec{D}_t$ reflect how users understand tag t and, if it changes, how that understanding changes through time.

To study the stream of documents $\vec{D}_t$ we define a sliding window covering l successive documents (Guha et al., 2003) that comprise a partial document collection under tag t. Typically, that window shifts by one document at a time, i.e. the least recent document within the sliding window is forgotten when a new tagging event is recorded for tag t. But such a fine-grained analysis is impractical for our purposes, because tag semantics do not change by one assignment of a tag to a single document. We rather slide the window by l_{new} documents, i.e. the window slides to a new position after l_{new} new documents have been annotated with tag t. Hence, the sliding window at position i covers the following partial sequence of documents $\vec{D}_t^i = \langle d_{r(i)}, \ldots, d_{r(i)+l} \rangle$ of $\vec{D}_t$ with $r(i) = 1 + (i-1)l_{new}$. Figure 1 shows an example with 22 tagging events. The sliding window covers $l = 7$ documents and it slides by $l_{new} = 5$ documents. The figure depicts four sequential positions of that sliding window, each covering a certain subsequence $\vec{D}_t^1, \vec{D}_t^2, \vec{D}_t^3$ and $\vec{D}_t^4$ of the stream $\vec{D}_t$ of documents.

As a result of applying the sliding window to the stream of documents under tag t, we get a sequence of $\bar{N}_t$ subsequences of document IDs $\langle \vec{D}_t^1, \ldots, \vec{D}_t^{\bar{N}_t} \rangle$.

4 LEARNING DOCUMENT PROTOTYPES

We use AdaptivePLSA (Gohr et al., 2009), which is an extension of probabilistic latent semantic analysis (PLSA) for topic modeling over streaming document collections. We review PLSA and briefly explain how AdaptivePLSA evolves a sequence of PLSA models from which we extract topics over times. These are used to derive document prototypes for summarizing of evolving document contents over time.

4.1 Topic Modeling

We use PLSA to extract K hidden topics for each sequence of documents $\vec{D}_t^i$ under tag t. We denote the set of all words (vocabulary) seen in documents of $\vec{D}_t^i$ as V_t^i. Topics are denoted by the unobserved variable z which takes values $1 \leq z \leq K$. Each topic is represented by a multinomial distribution $p(w|z)$ over word IDs $1 \leq w \leq |V_t^i|$. The data D_t^i used to learn topics is a set of triples (d, w, n) meaning that word with ID w is seen $n > 0$ times in document with ID d. If document d occurs m times in $\vec{D}_t^i$ then we increase all corresponding word counts by the factor m: $(d, w, n * m)$.

PLSA models word distributions of documents as mixtures of the determined topics: $p(w|d) = \sum_{z=1}^{K} p(z|d)p(w|z)$. The probabilities $p(z|d)$ are mixture weights for document d.

The parameters of a PLSA model ζ_t^i are:

- **document probabilities** which form a vector $\vec{\delta}$ with elements $\delta_d = p(d), d \in \vec{D}_t^i$,
- **mixture weights** which form a matrix $\vec{\theta}$ with elements $\vec{\theta}_d = (\theta_{1d}, \ldots, \theta_{Kd})$ and $\theta_{kd} = p(z = k|d), d \in \vec{D}_t^i$, $1 \leq k \leq K$, and

- **topics** which form a second matrix $\vec{\omega}$ with elements
  $\omega_{kw} = p(w|z=k), 1 \leq w \leq |V_t^i|, 1 \leq k \leq K$.

Usually K is much smaller than the number of documents in $\vec{D}_t^i$ but greater than one to capture the dominant word correlations.

Because PLSA is a probabilistic model it defines the probability of some data given the trained model:

$$\begin{aligned} p(D_t^i|\zeta_t^i) &= \prod_{j=1}^{|D_t^i|} p((d,w)_j|\zeta_t^i)^{n_i} \\ p((d,w)_j|\zeta_t^i) &= p(d_j)p(w_j|d_j) \\ &= p(d_j)\sum_{z=1}^{K} p(w_j,z|d_j) \\ &= p(d_j)\sum_{z=1}^{K} p(w_j|z)p(z|d_j) \end{aligned}$$

The last line follows because words and documents are assumed to be conditionally independent if the hidden topic from which the word comes is known.

Informally, estimating the parameters of a PLSA model for some given data means to find topics and mixture weights such that the word distributions $p(w|d)$ for the training documents are as best as possible approximated. The expectation maximization algorithm (EM) (Dempster et al., 1977) is used for parameter estimated because it allows to estimate model parameters even in presence of hidden (unobserved) variables like variable z.

4.2 Topic Modeling over a Stream of Documents

So far we had a closer look on how topics are learned for each subsequence of documents $\vec{D}_t^i$ under a tag t. The contents of documents of these subsequences $\langle \vec{D}_t^1, \ldots, \vec{D}_t^{\bar{N}_t} \rangle$ indicate emerging or abandoned meanings of tag t. Studying document prototypes derived from topics over time might reveal such changing meanings of the tag.

A problem of modeling resources under a tag over time is that these may introduce new words. Consequently, AdaptivePLSA evolves PLSA models under a tag over time by taking account for this volatility in vocabulary of the growing document collection. In addition, because AdaptivePLSA evolves the later PLSA model from the former one, the k^{th} topic of the later model evolves from the k^{th} topic of the former model.

The sliding window at a certain position i covers the sequence $\vec{D}_t^i$ of documents which partially overlaps with $\vec{D}_t^{i+1}$ (see Figure 1). We denote by ${}^{new}\vec{D}_t^i$ the latest l_{new} documents of $\vec{D}_t^i$ covered by the i^{th} sliding window. AdaptivePLSA adaptively learns a sequence of PLSA models $\zeta_t^1, \ldots, \zeta_t^{\bar{N}_t}$ for the stream of documents under tag t; i.e. it evolves the later PLSA models from the former ones. The PLSA model learned for the i^{th} position of the window ($\vec{D}_t^i$) is denoted by ζ_t^i.

To evolve model ζ_t^i into ζ_t^{i+1} AdaptivePLSA adapts ζ_t^i first to new documents and then to new words by five steps.

Estimating Mixture Weights for New Documents. Mixture weights of new documents ${}^{new}\vec{D}_t^{i+1}$ are estimated by folding-in (Hofmann, 2001) these documents into ζ_t^i. Therefore, topics $p(w|z)$ $1 \leq z \leq K$ are fixed and the EM algorithm estimates mixture weights $p(z|d)$ for new documents. Only that part of the new documents is considered that consists of words which are already known by model ζ_t^i.

Removing Mixture Weights of Old Documents. The first $l - l_{new}$ documents of window i are "outdated". Their mixture weights $\vec{\theta}_d$ are removed from $\vec{\theta}$.

Integrating New Words. New words are folded-in, which is not straight forward because words are connected by the word distributions $p(w|z)$. To allow folding-in of new words AdaptivePLSA converts the current model

$$p(w,d) = p(d)\sum_{z=1}^{K} p(w|z)p(z|d) \quad (1)$$

into the equivalent model by Bayesian Calculus

$$p(d,w) = p(w)\sum_{z=1}^{K} p(d|z)p(z|w)$$

Informally, documents and words have changed their roles. Thus, folding-in words is done analogously as we have previously folded-in new documents. We fix parameters $p(d|z)$ and use the EM algorithm to estimate $p(z|w)$ for the new words. The EM uses data which consists of occurrences of these words in the new documents.

Removing Old Words. Words that are not seen in documents of $\vec{D}_t^{i+1}$ are removed by deleting the corresponding parameters $p(z|w)$ for all $1 \leq z \leq K$.

Consolidation. To allow adaption to new words and new documents AdaptivePLSA converts back the PLSA model (Eq. 1) and runs the EM algorithm a few iterations using all data D_t^{i+1}. Thereby, it adapts mixture weights and topics.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

5 VISUALIZING DOCUMENT PROTOTYPES

The goal of our proposed visualization technique, called Topic Table, is to present K comprehensible document prototypes and their evolution over time. The document prototypes are derived from topics of PLSA models $\zeta_t^1, \ldots, \zeta_t^{\bar{N}_t}$ – each model learns K topics– learned for documents $\langle \vec{D}_t^1, \ldots, \vec{D}_t^{\bar{N}_t} \rangle$.

Topic Table arranges pieces of information in a table. For tag t, Topic Table has K rows and $\bar{N}_t$ columns. The cell (k,i) in row k and column i corresponds to the k^{th} prototype derived from the k^{th} topic of the PLSA model ζ_t^i. Hence, the rows correspond to document prototypes and the columns correspond to snapshots of these prototypes over time. By arranging the k^{th} document prototypes in one row, Topic Table establishes a correspondence among them. This correspondence stems from the fact that AdaptivePLSA evolves model ζ_t^{i+1} from the former model ζ_t^i for all $1 \leq i \leq \bar{N}_t - 1$. Hence, the k^{th} topic of model ζ_t^{i+1} evolved from the k^{th} topic of model ζ_t^i. Inspecting the sequence of the prototypes along the k^{th} row allows to better deduce how they change over time.

The Topic Table arranges three pieces of information for each document prototype and time point in different layers. From background to foreground, these pieces are i) how fast does a topic change between successive time periods, ii) how prominent a topic and the derived prototype are in documents under tag t during a certain period of time, and iii) the corresponding document prototypes. Figure 2 depicts how these pieces of information are visually presented by Topic Table.

First, we must visually depict how the learned topics change between two successive time points. To achieve this, we propose the metaphor of a *river* that "flows through time" and associate each evolving topic with a river. Narrow parts of the river represent watergates that strongly separate what comes before and what afterward. These watergates indicate time points at which the corresponding topic changes much. Topic Table visualizes the rivers as gray straps along each row, which correspond to the evolution of one topic over time. The width of each river changes between successive cells to indicate watergates. Successive cells, say (k,i) and $(k,i+1)$, correspond to the k^{th} topic of model ζ_t^i and ζ_t^{i+1}, respectively. These topics are multinomial distributions, which can be represented by two vectors $\vec{\omega}_k^i$ and $\vec{\omega}_k^{i+1}$. The entries are probabilities of words of the respective vocabularies of documents of $\vec{D}_t^i$ and $\vec{D}_t^{i+1}$. The more similar these two vectors are the more stable the corresponding topic is. We use the cosine similarity which is equal to 1 if both vectors point into the same direction and equal to 0 if the vectors are orthogonal to each other. Hence, the width of the river at the border between cells (k,i) and $(k,i+1)$ is proportional to the determined similarity between $\vec{\omega}_k^i$ and $\vec{\omega}_k^{i+1}$ such that when the similarity is equal to one the width of the river would be equal to the height of the cells. Vocabularies of $\vec{D}_t^i$ and $\vec{D}_t^{i+1}$ are likely to be different. To compute cosine similarity between the vectors $\vec{\omega}_k^i$ and $\vec{\omega}_k^{i+1}$ which might be defined in different spaces, we embed them into the joint space defined by the union of both vocabularies.

Another useful information is the relative strength of the learned topics and corresponding document prototypes in the data over time. This kind of information is helpful in two respects. First, a user might want to study only the strongest document prototypes at each time point. Second, a user wants to inspect at a glance temporal patterns of strong prototypes; how the strength of them changes over time. Each PLSA model allows to derive the probabilities of each extracted topic $p(\text{z}{=}k)$. All these probabilities sum to one $1 = \sum_{k=1}^{K} p(\text{z}{=}k)$ for each studied time period. A large probability indicates a topic that is prevalent in the data. Topic Table visualizes these probabilities by circles in the center of each cell (see (c) in Figure 2); The probability is mapped to the area of the circle. We follow (W. S. Cleveland, 1994) and map the strength of a topic to the circle area in a nonlinear way to enhance human perception of differences in the quantity; The radius of the circle visualizing the probability of a topic k is equal to $\frac{p(z=k)^{5/7}}{\sqrt{2\pi}}$.The circles are depicted in the background of the cell on top of the background river. Studying all circles of a column top-down gives a fast impression about what topics are the most dominant ones at a certain time period. Inspecting the circles along a row allows to deduce how the relative strength of the corresponding topic changes over time.

Last, Topic Table shows the document prototypes for each topic consisting of the most likely words. Topic Table lists these words in the foreground of the corresponding cells. The number of words that constitute a document prototype is not fixed. Common choices are ten to twenty words. An experienced user may need only ten-word prototypes ($N_{top} = 10$) to deduce what they are about. New users or users who study resources under unknown tags might need more words. Consequently, Topic Table lets the user decide how many words should constitute the document prototypes. To assist users who want to find new words Topic Table highlights words of the k^{th} prototypes at time point $i+1$ that are not part of the k^{th} prototype

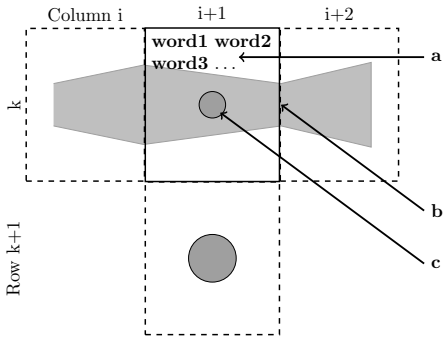

Figure 2: The cell in row k and column $(i+1)$ of Topic Table corresponds to the k^{th} topic extracted from the documents $\vec{D}_t^{i+1}$ under tag t. Features of Topic Table are (a) the N_{top} most likely words per topic that define the corresponding document prototype (bold face words are new and not part of the previous prototype), and (b) the river in the background of each row has a width at each border between the $(i+1)^{\text{th}}$ and $(i+2)^{\text{th}}$ that is proportional to the similarity between the k^{th} topic at $(i+1)^{\text{th}}$ and $(i+2)^{\text{th}}$ position of the sliding window, and (c) the radius of the background circle is non-linearly proportional to the probability/strength of the corresponding topic/derived prototype.

at time point i.

To enhance perception which visual elements of Topic Table belong to the same layer, Topic Table uses different gray shades. All background rivers are shown in light gray. Light gray elements are often strongly assigned to the background. Because the rivers are drawn along a row they combine cells of one row and thereby strengthen the perception of rows and hence conveys the evolution of the document prototypes through time. The circles belong to another layer on top of the background layer. Consequently, they are all drawn in a darker gray shade. The darker the elements, the more important they are assumed to be. Because the circles are positioned in the center of the cells they enhance the perception of the structure of Topic Table. Last, the document prototypes are listed in the foreground. Words are written in black what enhances perception as important foreground elements.

By displaying the different kinds of information in layers on top of each other, Topic Table uses the canvas efficiently. In addition, because Topic Table uses always one cell per document prototype less dominant prototypes are visually retained and not suppressed.

6 Bibsonomy CASE STUDY

To show how our visualization technique helps in clarifying the semantics of ambiguous tags, we run experiments on the Bibsonomy social platform.

6.1 Data Preparation and Parameter Setting

Resources in Bibsonomy are bibliographic entries in Bibtex format which were contributed by the users between 2005-12-31 and 2008-12-31. We use the cleaned dump of the Bibsonomy [5]. We enriched some Bibsonomy entries by retrieving abstracts from the ACM Digital Library[6]. Entries are omitted when they contain an insufficient abstract being shorter than 400 characters. German and French abstracts were pruned by a simple heuristics that checks for German articles[7] and "sociaux". The remaining English abstracts, which we call documents, are subjected to standard preprocessing techniques, i.e. stopword removal and Porter stemming.

Three parameters influence the visualization and AdaptivePLSA. First, the length l of the sliding window determines the number of documents a particular PLSA model is trained on. It also specifies implicitly with respect to what time scale a PLSA model is computed, meaning the difference in time between the last and least covered document. The parameter l might be adapted to the amount of available documents under a tag. The parameter l_{new} controls how fast the sliding window moves over the document stream under a tag. Without making further assumptions it can be meaningfully varied between 1 and l. We set $l_{new} = 0.75 \cdot l$ to force some overlap while analyzing the streams under tags in a rough manner. The number of hidden topics K learned by a PLSA model affects the roughness of the summary of resources. A reasonable choice is $K \ll l$ if a rough summary of the contents of documents is desired. Consequently, we set K equal to 5.

6.2 Topic Table for Tag *network*

The tag *network(s)*, having many meanings as we will see, stands for the two tags *network* and *networks*. That tag was assigned to 1218 documents from 2006-01-24 until 2008-12-27. The sliding window covers 350 documents and moves by 75% (260 documents) of its length. This setting results in four positions of the sliding window covering 350 documents. Hence, Topic Table summarizes the contents of documents covered by the four window positions in four columns as shown in Figure 3.

[5] www.kde.cs.uni-kassel.de/ws/dc09/dataset It was part of the data mining contest of the ECML/PKDD conference 2009.

[6] www.acm.org

[7] der, die, das, dieser, diese, dieses, ein, eine, einer

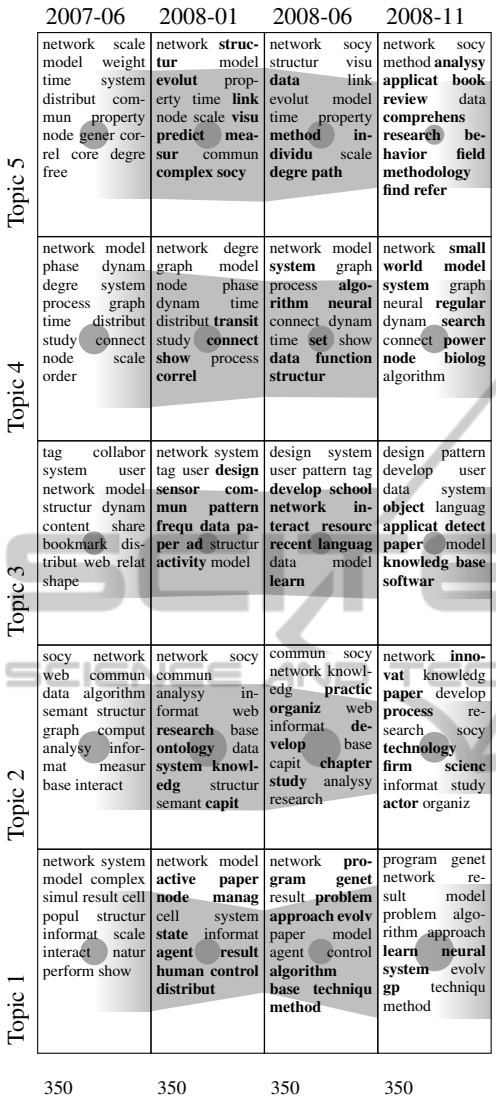

Figure 3: Topic Table for tag *network(s)*. The sliding window covers 350 documents. The window moves by 75% of its length (262 documents). Document prototypes consist of the top 15 most likely words (stemmed by Porter stemmer) for each extracted topic. Time stamps at the top are those of the last document covered by the sliding windows. Numbers at the bottom indicate the number of summarized documents.

The bottom row shows prototypes over time derived from one of the five topics under tag *network*. At the first time point, 2007-06, it is associated with the stemmed words *cell*, *natur* and *simul* which might stand for research about biological networks or simulation of biological processes using networks. At the next time point that topic changes and the prototype consists of words like *genet* and *program* which might stand for aspects of genetic programming. One time point later, the prototype uncovers aspects of neural networks and learning approaches using genetic programming.

The second derived prototype emphasizes until 2008-06 aspects of social networks, their analysis, ontologies and usage of social networks to organize data. Later, the role of networking among firms but also among research communities to enhance the process of innovation seems to emerge. Perception of that change is visually supported by the river in the background of the second row: Its width decreases at the transition from 2008-06 to 2008-11.

The third prototype focuses first on collaborative tagging systems used to manage shared and distributed resources, e.g. bookmarks. At 2008-01 that prototype gets more diverse: It is enriched by aspects of sensor and communication networks, their design and analysis of frequent patters in those networks. Then, at 2008-06, stemmed words like *school*, *languag* and *learn* give hints that some content under tag *network(s)* is about networks of schools to improve learning of pupils. At 2008-11 words *develop*, *applicat* and *softwar* bring aspects of software development, e.g. for network applications into play. The liveliness of the underlying topic is also indicated by the river in the background of the third row: It is relatively narrow over the whole period of time.

Topic Table lists for the fourth prototype at the beginning words *dynam*, *degre*, *process* and *time* which might stand for the analysis of network features like node degree. Further inspecting the fourth row, we see that the prototype seems to evolve through 2008-11 toward neural networks, while the word *biolog* may refer to biological networks or to life-inspired networks.

Prototype five, at the top of Figure 3, is at first about scale-free networks. At the next time point we see that the aspect of social networks arises, associated with the word *visu* that indicates documents on network visualization. At the last time point this topic seems to have drifted toward literature on networks (e.g. books and reviews), while the words *behavior* and *individu* may refer to individual behavior in social networks.

At two points in time the background rivers are especially narrow. The river corresponding to the first topic has a watergate at the transition from the second to the third cell. We indeed find that the corresponding derived prototype changes drastically; The accentuation on biological networks disappears, and genetic programming appears. Second, the river that indicates how the second topic changes over time is especially narrow at the transition from the third to the last cell. Again, at this transition we find an obvious change from accentuation on social networks

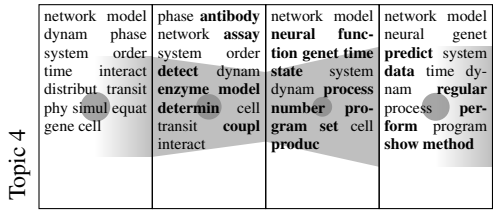

Figure 4: Part of Topic Table for tag *network(s)*. Same parameter setting as for Figure 3. 25 documents annotated with tag *immunoassay* have been mixed in so that they will be covered by sliding window at its second position (second cell).

to the issue of networking among (or in) technology firms.

Inspecting the circles along the Topic Table, we find that the second topic, and hence its derived prototype, dominates especially during the time period covered by the second and third position of the sliding window, meaning that social networks are a prevalent topic under tag *network(s)*. The dominance of the second topic disappears in the last period of time while the first topic becomes the most dominant one indicating that neural networks are another important topic under tag *network(s)*.

6.3 Effectiveness

To show that reported prototypes are not only artifacts but indeed summarize contents of documents under tags we did the following experiment. We added 25 documents annotated with the tag *immunoassay* to the stream of documents under the tag *network(s)*. These documents have been added such that all of them are covered by the sliding window at its second position. Because this sliding window covers 350 documents in total, documents annotated with *immunoassay* are a fraction of only about 7% of all covered documents. Figure 4 shows the fourth topic of Topic Table, which is the only one that has changed dramatically. We find three words *antibody*, *assay* and *enzyme* among listed words in the second cell that corresponds to the time period with alien documents. The emergence of three stemmed words corresponding to documents annotated with the tag *immunoassay* within the 15 most likely words demonstrates the effectiveness of AdaptivePLSA and the proposed visualization technique to summarize document contents and capture their evolution over time.

7 CONCLUSIONS

We propose Topic Table, a new visualization technique for studying the evolution of contents in a stream of documents. Topic Table visualizes document prototypes learned in an unsupervised manner by topic models like PLSA.

We apply Topic Table and PLSA to analyze the document content over time under a tag of the collaborative tagging system Bibsonomy that aims at sharing bibliographic entries. By inspecting the ambiguous tag *network(s)* and by finding a bunch of of themes tag *network(s)* is associated with (e.g. social, neural and biological networks), we show that Topic Table summarizes document contents over time in a clear and apprehensive fashion. With respect to time, Topic Table indicates that social networks are prevalent from 2007-06 to 2008-06 and that later neural networks become the strongest single aspect. We demonstrate the effectiveness of our approach by adding some alien documents to the documents under tag *network(s)*. Re-learning topics over time and visualizing them by Topic Table, we indeed find one prototype that indicates the existence of the alien documents although their minor abundance.

Because of Topic Table's general applicability to visualize topics over time learned by any available topic modeling method, we believe Topic Table has the potential to become a general tool for visualizing and summarizing document contents changing over time.

REFERENCES

AlSumait, L., Barbara, D., and Domeniconi, C. (2008). On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In *ICDM*.

Blei, D. and Lafferty, J. (2006). Dynamic topic models. In *ICML*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Boyd-Graber, J., Chang, J., Gerrish, S., Wang, C., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*.

Chou, T.-C. and Chen, M. C. (2008). Using incremental PLSI for threshold-resilient online event analysis. *IEEE Trans. on Knowl. and Data Eng.*, 20(3):289–299.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Gohr, A., Hinneburg, A., Schult, R., and Spiliopoulou, M. (2009). Topic evolution in a stream of documents. In *SIAM Data Mining Conf. (SDM'09)*, pages 378–385, Reno, CA.

Goyal, A., Bonchi, F., and Lakshmanan, L. V. (2008). Discovering leaders from community actions. In *CIKM'08*, pages 499–508, Napa Valley, CA, USA. ACM.

Guha, S., Meyerson, A., Mishra, N., Motwani, R., and O'Callaghan, L. (2003). Clustering data streams: Theory and practice. *IEEE Trans. of Knowlende and Data Eng.*, 15(3):515–528.

Havre, S., Hetzler, E., Whitney, P., and Nowell, L. (2002). Themeriver: Visualizing thematic changed in large document collections. *IEEE Trans. Visualization and Computer Graphics*, 8(1):9–20.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

Mei, Q., Shen, X., and Zhai, C. (2007). Automatic labeling of multinomial topic models. In *KDD*, pages 490–499.

Mei, Q. and Zhai, C. (2005). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *SIGKDD*, pages 198–207, New York, NY, USA. ACM.

Moxley, E., Kleban, J., Xu, J., and Manjunath, B. S. (2009). Not all tags are created equal: learning flickr tag semantics for global annotation. In *ICME'09: Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, pages 1452–1455, Piscataway, NJ, USA. IEEE Press.

Suchanek, F. M., Vojnovic, M., and Gunawardena, D. (2008). Social tags: Meanings and suggestions. In *CIKM'08*, pages 223–232, Napa Valley, CA, USA. ACM.

W. S. Cleveland (1985, 1994). *The Elements of Graphing Data*. Hobart Press, Summit, New Jersey, U.S.A.

Wang, X. and McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In *SIGKDD*, pages 424–433. ACM.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS