Automatically Tagging Online Community Content using Probabilistic Topic Models

M. Döhring\textsuperscript{1,2}, A. Berlea\textsuperscript{1}, U. Störl\textsuperscript{2}

doe@stud.fbi.h-da.de, u.stoerl@fbi.h-da.de, alexandru.berlea@sap.com

\textsuperscript{1} SAP Research, \textsuperscript{2} University of Applied Sciences Darmstadt

1 Motivation and Scope

The continuous growth of information available on the WWW leads to an increase in information overflow in current times. If a search query is not very precisely specified, one can expect hundreds or thousands of potentially relevant search results. The assignment and use of keyphrases, which are relatively small sets of terms reflecting the main topics treated within a document is known of being suitable to mitigate this problem. In collaborative online platforms, people assign their own keyphrases to a variety of items, predominantly web pages, but also to images, video clips, restaurant locations and many more. A user is thereby able to see what other keyphrases users already have assigned to the items. This collaborative keyphrase indexing process is called tagging [Med09].

Companies currently carefully inch their way forward concerning the provision of Web 2.0 features to their customers and a majority of online communities in this sector is still in its formation phase. When launching a community, it is wise to provide some initial content and meta-data in the community to encourage user participation and to attract new users. More precisely, when providing a set of documents within the community, it is desirable to also provide some initial tags that can be used to find information and be reused for new documents created by the community members. The crucial problem at this point is, that the manual creation of a controlled vocabulary is usually complex and expensive [Qui05]. This work therefore sketches a completely unsupervised approach for assigning tags to documents from a collection, like e.g. software documentation pages. The tags should be chosen by their ability to best reflect the topic mixture a document deals with.
### 2 Outline of Topic-Based Auto-Tagging

The main problems with existing even state-of-the-art unsupervised approaches for keyphrase extraction or auto-tagging are that they either only observe single documents (e.g. [MT04], [GGL09]) or take only single word tags into account (e.g. [DWPP08]). Consequently, it is sometimes hard to detect multi-word terms like `CLIENT_CERTIFICATE` which have a low occurrence count but a high relevancy for a document dealing with a topic mixture of portal landscapes and user authentication as tags for the document.

For finding an explicit representation of topics corresponding to human intuition in a document collection, [SSRZG04] have proposed the probabilistic topic model (PTM) approach. PTM assume an inherent latent semantic structure in a document corpus which the approach tries to reveal by examining the co-occurrences of words in documents. On can imagine a topic as pattern of word co-occurrences that can be found usually in several documents within the corpus. By dint of the occurrence probabilities inferred for PTM, one can assign relevancy figures for each word-topic and each topic-document relation.

Two example topics detected in SAP NetWeaver Composition Environment 7.1 technical document set together with their top 8 words in order of decreasing relevance are presented in Table 1. Just by examining the two word sets, the reader can conclude that topic A deals with databases, while topic B is about synchronization.

<table>
<thead>
<tr>
<th>Topic</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>database</td>
<td>sql</td>
<td>oracle</td>
<td>maxdb</td>
<td>db</td>
<td>connection</td>
<td>table</td>
<td>jdbc</td>
</tr>
<tr>
<td>B</td>
<td>lock</td>
<td>transaction</td>
<td>enqueue</td>
<td>exclusive</td>
<td>release</td>
<td>entry</td>
<td>rollback</td>
<td>owner</td>
</tr>
</tbody>
</table>

The general idea of looking on topics as probability distributions over single words has been extended by [WMW07] for multi-word terms. One main contribution of our work is the sound formulation of a topic-based relevancy measure for (multi-word) tags within document collection and the proper definition of its alignment within a pre- and post-processing pipeline as depicted in Figure 1. As a proof of concept, this pipeline has been implemented using LinguistX Platform\(^1\) for tokenization, POS-tagging and lemmatization and Mallet [McC02] for model inference.

The pipeline takes arbitrary text or HTML documents and first removes irrelevant content sections e.g. by dint of predefined XPath patterns. Next it applies standard natural language preprocessing steps removing tendentially irrelevant word tokens and collapsing similar word forms to their common morphological root. It then executes the inference procedure described in [WMW07], which is a kind of Markov-Chain Monte-Carlo. All proper-noun single words, multi-words which are considered to belong together in most of the states of the Markov-Chain as well as additional terms (e.g. a whitelist of product names) are extracted as tag candidates. Based on the topic model and an optionally available thesaurus, our approach is also capable of deciding e.g. whether to expand abbreviations like UME to unplanned material exchange or user management engine. After

\(^{1}\)http://www.inxightfedsys.com/products/sdk/1x
the top-ranked candidates according to our proposed measure have been determined and have gone through a post-processing step (e.g. filtering tags not assigned to a minimum of other documents in the collection), the tags are attached to their corresponding documents.

3 Results of the User Evaluation

The implementation described in the previous section has been applied to 12801 online documentation pages of SAP SAP NetWeaver Composition Environment 7.1. The 10 randomly drawn representative documents together with their automatically assigned tags were presented to two user groups via a web interface. One group consisted of 30 SAP employees, one of 11 students in computer science. All participants were presented the automatically assigned tags for rating as well as eventually tags from other users. Moreover, missing tags could be added per document. With the obtained indexer-consistency and tag suitability measures obtained from the ratings, the following findings can be supported:

- Tagging in technical documents is a quite subjective issue, as users partly tend to disagree on what is the optimal tag set for a document.
- The tag sets delivered by our approach are at least comparably accepted by the participants than those tags added by other participants.
- Our approach tendentially produces tags that are better accepted by participants when rated in isolation than those tags added by other participants (e.g. highly important for tag-clouds).

Moreover, on public gold-standards ([NK07], [Hul04]) for keyphrase extraction, we achieved comparable or better performance than [ZG09] reported for state-of-the-art procedures.
4 Future Work

Our approach could benefit from enhancing the relevancy measure by including reference corpus statistics e.g. from Wikipedia. Furthermore, we did not fully exploit the richness of HTML information (like bold tags). Regarding the growth degree of online content, also the scalability of the topic model inference algorithm should be studied further.

References


