Term extraction for user profiling: evaluation by the user

Suzan Verberne[1], Maya Sappelli[1,2], Wessel Kraaij[1,2]

[1] Institute for Computing and Information Sciences, Radboud University Nijmegen
[2] TNO, Delft
s.verberne@cs.ru.nl

Abstract. We compared three term scoring methods in their ability to extract descriptive terms from a knowledge worker's document collection. We compared the methods in two different evaluation scenarios, both from the perspective of the user: a per-term evaluation, and a holistic (term cloud) evaluation. We found that users tend to prefer a term scoring method that gives a higher score to multi-word terms than to single-word terms. In addition, users are not always consistent in their judgements of term profiles, if they are presented in different forms (as list or as cloud).

1 Introduction

In our project we aim to develop smart tools that support knowledge workers in their daily life. One of our objectives is a tool for personalized information filtering. We focus on two information filtering tasks: e-mail organization (which messages are important, which messages are related to a specific project) and professional search. In order to help the user to select relevant and important information in the large body of incoming e-mails and online search results, we need to create a model of the user. In the current work, we focus on the content-based part of the user profile: the user-specific terminology.

We aim to develop a user term profile that serves two purposes: (1) it will be used by our filtering tool for estimating the relevance of incoming information, and (2) it should give the user insight in his or her profile: which terminology is important in which context, and which terminology is shared with co-workers? Thus, the user profile should not only be effective in a system context but also valued by the user.

In the current paper, we evaluate three term scoring methods for the purpose of user profiling. We compared the methods in two different evaluation scenarios, both from the perspective of the user: a per-term evaluation, and a holistic (term cloud) evaluation. In Section 2 we describe three methods for collecting the descriptive terms from a user's self-authored document collection, and the evaluation setup. In Section 3, we present the results from the three methods and compare the two evaluation scenarios. Section 4 describes our conclusions and plans for future work.

2 Methodology

The input for our term extraction technology is a document collection provided by the user. First, the document collection is preprocessed: Each document is converted to plain text and the documents are split in sentences. Then candidate terms are extracted: Given a document collection, we consider as candidate terms all occurring n-grams (sequences of n words) that contain no stop words and no numbers. We used $n = [1, 2, 3]$ for the candidate terms. All candidate terms are saved with their term counts.

2.1 Term scoring methods

Until now, term scoring methods have mainly been evaluated in the context of Information Retrieval [8] and text classification [5, 10]. In Information Retrieval, term weighting is used to estimate the relevance of a document to a query. Therefore, term weighting in IR is generally based on TF-IDF (term frequency–inverse document frequency): a term is weighted by its frequency in the document, and by the number of documents in the corpus in which it occurs. Terms that occur in more documents are less informative for the documents in which they occur. In text classification, term weighting is used to select the terms that are the most informative for a specific category. Chi-square for example measures the lack of independence between a term and a category and Information Gain measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document [11].

The goal of term scoring for user profiling is to find the terms that are the most descriptive for a user's corpus. A way to select informative terms that are distinctive for the user's corpus compared to general English, we use a background corpus. We chose the Corpus of Contemporary American English as background corpus, which is free to use and is easy to process because the developers provide a word frequency list and n-gram frequency lists. We implemented three different term scoring functions from the literature:

1. Parsimonuous language model based (PLM): A method based on [2] where term frequency in the personal collection is weighted with the frequency of the term in the background corpus.

$$e_t = tf(t, D) * \frac{\lambda P(t|D)}{(1-\lambda)P(t|C) + \lambda P(t|D)} \tag{1}$$

$$P(t|D) = \frac{e_t}{\sum\limits_t e_t} \tag{2}$$

   Here, $P(t|D)$ is the probability of the term t in the personal document collection, $P(t|C)$ is the probability of the term in the background corpus and λ is a parameter that determines the strength of the contrast between foreground and background probabilities.

2. Cooccurence based (CB): A method based on [6] where term relevance is determined by the distribution of co-occurences of the term with frequent terms in the collection. The rationale is of this method is that no background corpus is needed because the most frequent terms from the foreground collection serve as background corpus.

   $$\chi^2(t) = \sum_{g \in G} \frac{freq(t,g) - n_t p_g)^2}{n_w p_g} \tag{3}$$

   $$\chi^{'2}(t) = \chi^2(t) - \max_{g \in G}\{\frac{freq(t,g) - n_t p_g)^2}{n_t p_g}\} \tag{4}$$

   Here, G is the set of frequent terms (the size of which is determined by the parameter $topfreq$), $freq(t,g)$ is the co-occurrence frequency (in sentences) of t and g, n_t is the total number of co-occurrences of term t and G, and p_g is the expected probability of g.
3. Kullback-Leibler divergence for informativeness and phraseness (KLIP): A method based on [9] where the term relevance is based on the expected loss between two language models, measured with point-wise Kullback-Leibler divergence:

   $$P(p||q) = \sum_x p(x) log \frac{p(x)}{q(x)} \tag{5}$$

   Tomokiyo and Hurst propose to mix two models for term scoring: phraseness (how tight are the words in the sequence) and informativeness (how informative is the term for the foreground corpus). The parameter γ is the weight of the informativeness score relative to the phraseness score.

The result of each of the term scoring methods is a list of terms for a document collection, with scores. We used the following parameter settings in our experiments: $\lambda = 0.5$ in the PLM method, $topfreq = 10$ in the CB method. In the KLIP method, we decided to give more weight to the phraseness component than to the informativeness component, because this is the only method that has a phraseness component. We set gamma to 0.1, which leads KLIP to generate more multi-word terms than the other methods.[3] We note here that the parameters should be optimized in future work.

2.2 Evaluation set-up

We asked five colleagues to provide us with a collection of at least 20 documents that are representative for their work. On average, we received 22 English-language documents per user (mainly scientific articles) with an average total of around 537.000 words per collecton. For each of these document collections, we generated three lists with 300 terms each using the PLM, CB and KLIP methods. Then we created a pool of terms per collection by first normalizing the

[3] In [9], informativeness and phraseness are weighted equally.

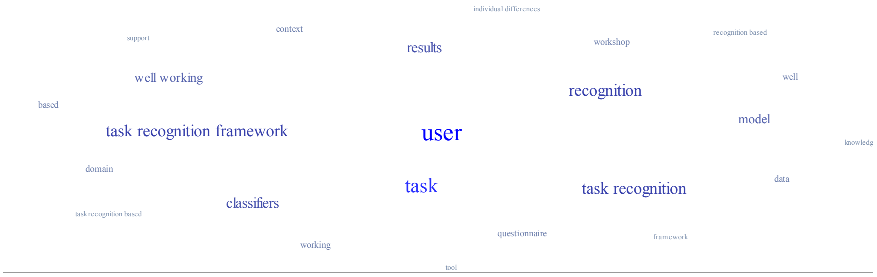

Fig. 1. Example of a tag cloud as it was shown to the user.

scores in each of the three lists relative to the maximum and minimum scores. We then calculated for each term the average of the three normalized scores. We ordered the terms by the combined scores and extracted the top-150. These terms were judged in alphabetical order by the owners of the document collections. We asked them to indicate which of the terms are relevant for their work. There was a large deviation in how many terms were judged as relevant by the users (between 24% and 51%), but on average, around one third of the generated terms (36%) was perceived as relevant.

In a second experiment, we evaluated the terms using term clouds. Instead of evaluating terms one by one, the profiles extracted from the documents were evaluated as a whole. Kaptein et al. [3, 4] and Gottron [1] show that using a term cloud as method to summarize a document can help the user in determining the topic of the document. For each user's document collection, we generated term clouds using the three term scoring methods. We chose a term cloud visualization where the biggest term is in the center of the cloud and the 25 subsequent terms are added in a spiral form, ending with the smallest terms in the outer ring. An example is shown in figure 1. We showed the term clouds in random order to the owners of the document collections, and asked them to rank the three clouds from the best to the worst representation of their work. They were allowed to give the same rank to two clouds, if they judged them equal in quality.

3 Results

We ordered the term lists by term scores from high to low and then used the term assessments to evaluate the ranked term lists for the three scoring methods. As evaluation measure we used Average Precision [12]:

$$\text{Average Precision} = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{n_c}, \tag{6}$$

where $P(k)$ is the precision at rank k, n is the total number of terms in the list, n_c is the total number of relevant terms and $rel(k)$ is a function that equals 1 if the term at rank k is a relevant theme, and zero if it is not relevant. The results are presented in the upper half of Table 1. The lower half of the table shows the results for the ranking of the term clouds by the users.

Table 1. Results for the evaluation of the term lists and term clouds, per user A–E and overall. TF scores is a baseline ranking based on simple term frequency.

	A	B	C	D	E	Average	Stddev
% of pooled terms judged relevant	49%	30%	29%	51%	24%	36%	13%
	Average precision of ranked list						
TF scores	0.388	0.299	0.213	0.448	0.166	0.303	0.118
PLM scores ($\lambda = 0.5$)	0.407	0.312	0.221	0.461	0.177	0.316	0.120
CB scores ($toprank = 10$)	**0.424**	0.319	0.217	0.441	0.207	0.322	0.111
KLIP scores ($\gamma = 0.1$)	0.409	**0.438**	**0.409**	**0.599**	**0.293**	**0.430**	0.110
	Ranks of the term clouds. 1=best; 3=worst						
PLM scores ($\lambda = 0.5$)	2	2	**1**	2	2	1.8	0.4
CB scores ($toprank = 10$)	**1**	3	2	3	**1**	2	1.0
KLIP scores ($\gamma = 0.1$)	2	**1**	**1**	**1**	2	1.4	0.5

The table shows a large variation in the evaluation scores for the five knowledge workers. All three term extraction methods give better results than the plain TF scores. For all users except one, the KLIP method generates the best ranked list. As explained in Section 2, KLIP extracts more multi-word terms than the other two methods because it has a phraseness component. In fact, in the top-100 of most descriptive terms, KLIP has 64 multi-word terms on average, compared to 5 for Hiemstra and 4 for Matsuo. The finding that KLIP is judged the most positive suggests that users tend to find multi-word terms better descriptors of their work than single-word terms. The ranking of the term clouds for the three methods is significantly correlated to the ranking of the methods based on the Average Precision scores (Kendall $\tau = 0.67, P = 0.008$) but there are some differences. For example, to user E, the KLIP method generated the best ranking, but she judged the CB cloud as best visual reresentation of her work domain. This suggests that the visualisation of a term profile can play a role in how the user perceives the profile.

We also asked the users to label the terms that they judged as irrelevant with a reason why the term was not relevant. The results of this categorization are in Figure 2. The figure shows that there are no big differences between the types of irrelevant terms selected by the term scoring methods.

4 Conclusions and future work

We compared three term scoring methods in their ability to extract descriptive terms from a knowledge worker's document collection. On a small group of five users, we found that Kullback-Leibler divergence incorporating not only infor-

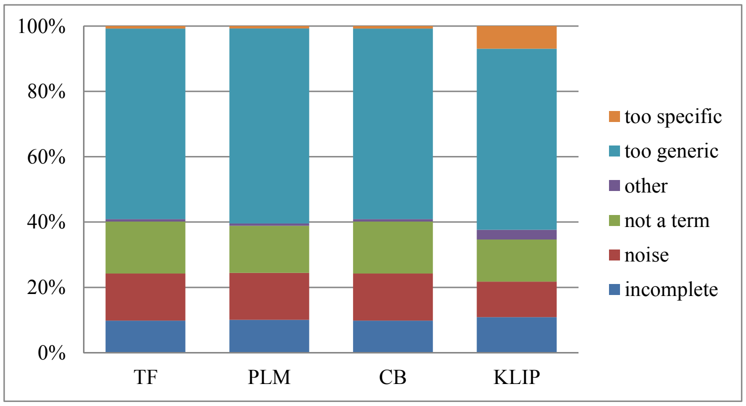

Fig. 2. Reasons that the users provided for irrelevant terms being irrelevant, per term scoring method. The counts have been summed over the users. 'Incomplete' denotes a partial term, e.g. care professional instead of health care professional; 'noise' are words in a different language, a PDF conversion error, parts of the document structure; 'not a term' are n-grams such as 'using' and 'million queries'.

mativeness but also phraseness of the terms gives the best results (Mean Average Precision is 0.43).

Since this work is still in an early stage, we can only draw preliminary conclusions. First, our results suggest that users tend to prefer a term scoring method that gives a higher score to multi-word terms than to single-word terms. It could be that multi-word terms are considered better descriptors because they are more specific than single-word terms. Second, users are not always consistent in their judgements of term profiles, if they are presented in different forms (as list or as cloud).

In the near future, we want to focus more on the best visualization of term profiles. For example, Rivadeneira et al. [7] found that tagclouds presented as an ordered list were easiest to comprehend. We will also study the possibilities of term clustering in order to visualize the multiple topics of projects that a knowledge worker is involved in, and investigate the differences between self-assessment of the profiles and judgments by colleagues. In addition, we will experiment with improving our term scoring methods by (1) optimization of the parameters λ (PLM), $topfreq$ (CB) and γ (KLIP), (2) finding an optimal combination of the three methods, (3) adding features to the terms such as position in the document, giving a higher preference to title words and (4) experimenting with a more specific background corpus. For example, in the Artificial Intelligence field, terms such as 'data' or 'user' are more frequent than in general English, but they might be considered too general to describe the work domain of one specific researcher in Artificial Intelligence.

References

1. Gottron, T.: Document word clouds: Visualising web documents as tag clouds to aid users in relevance decisions. In: Research and Advanced Technology for Digital Libraries, pp. 94–105. Springer (2009)
2. Hiemstra, D., Robertson, S., Zaragoza, H.: Parsimonious language models for information retrieval. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 178–185. ACM (2004)
3. Kaptein, R., Hiemstra, D., Kamps, J.: How different are language models andword clouds? In: Advances in Information Retrieval, pp. 556–568. Springer (2010)
4. Kaptein, R., Marx, M.: Focused retrieval and result aggregation with political data. Information retrieval 13(5), 412–433 (2010)
5. Lan, M., Tan, C.L., Su, J., Lu, Y.: Supervised and traditional term weighting methods for automatic text categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(4), 721–735 (2009)
6. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 13(01), 157–169 (2004)
7. Rivadeneira, A., Gruen, D.M., Muller, M.J., Millen, D.R.: Getting our head in the clouds: toward evaluation studies of tagclouds. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 995–998. ACM (2007)
8. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management 24(5), 513–523 (1988)
9. Tomokiyo, T., Hurst, M.: A language model approach to keyphrase extraction. In: Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18. pp. 33–40. Association for Computational Linguistics (2003)
10. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-. pp. 412–420. MORGAN KAUFMANN PUBLISHERS, INC. (1997)
11. Zheng, Z., Wu, X., Srihari, R.: Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter 6(1), 80–89 (2004)
12. Zhu, M.: Recall, precision and average precision. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo (2004), working paper