SEMANTIC IDENTIFICATION AND VISUALIZATION OF SIGNIFICANT WORDS WITHIN DOCUMENTS

Approach to Visualize Relevant Words within Documents to a Search Query by Word Similarity Computation

Karolis Kleiza, Patrick Klein and Klaus-Dieter Thoben
BIBA - Bremer Institut für Produktion und Logistik GmbH, Hochschulring 20, 28359 Bremen, Germany

Keywords: Similarity Computation, Word Similarity, Probabilistic Topic Model, Latent Dirichlet Allocation, Word Highlighting.

Abstract: This paper gives at first an introduction to similarity computation and text summarization of documents by usage of a probabilistic topic model, especially Latent Dirichlet Allocation (LDA). Afterwards it provides a discussion about the need of a better understanding for the reason and transparency at all for the end-user why documents with a computed similarity actually are similar to a given search query. The authors propose for that an approach to identify and highlight words with respect to their semantic relevance directly within documents and provide a theoretical background as well as an adequate visual assignment for that approach.

1 INTRODUCTION

Knowledge and information retrieval tasks have become a time-consuming challenge in nowadays business processes, such as the development process of a complex product or system. Even though contemporary search engines provide a quick access to company's specific databases, several studies has shown that employees spent up to 35 percent of their hours of work for searching for relevant documents (Bahrs et al., 2007). Especially an extraction of relevant and needed information within respective documents is not yet supported adequately. To optimally exploit the large amounts of engineering information as fast as possible it is often not sufficient to provide helpful documents. In practical use many documents, reports or guidelines cover diverse technological and methodical aspects and are not focused on a single topic. Hence a simple suggestion of such documents dissatisfies a user, since he obviously dislikes scanning the whole content just to identify only some interesting paragraphs.

In order to identify such paragraphs conventional keywords based document retrieval allows in subsequence a search within a document by usage of the same keywords. In contradiction to this the proposed search approach of semantic similarity between helpful documents and a given working context abstains from user based keywords thus making a subsequent internal search complicated or impossible. There is a need for an enhancement by providing functionalities to decide quickly, if a suggested document is appropriate for the actual work context, in terms of *why is the specific document highly ranked* and *which document parts are interesting*.

Hence the overarching vision addressed by the presented approach is a practical solution, where a result list of a (semantic) search is enhanced by a direct preview of respective documents. This preview has to come up with highlighted or marked up keywords and text passages in a way, that a user will be enabled to identify easily the most relevant sections with respect to his/hers actual working-context at a glance. Ideally by use of different colors or color intensities additional information, such as "importance to context" should be given to support a quick and easy interpretation of the given results.

2 BACKGROUND

Several methods for generating semantics out of a

Kleiza K., Klein P. and Thoben K..
SEMANTIC IDENTIFICATION AND VISUALIZATION OF SIGNIFICANT WORDS WITHIN DOCUMENTS - Approach to Visualize Relevant Words within Documents to a Search Query by Word Similarity Computation.
DOI: 10.5220/0003099004810486
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 481-486
ISBN: 978-989-8425-28-7

Table 1: Four selected topics with representative labels of the high lift device domain in the aviation industry.

Topic 4: "aero design"	Topic 7: "lessons learned"	Topic 13: "components"	Topic 14: "fly-by-wire"
aerodynamic	test	rib	flap
shape	stress	spar	airbus
design	structure	plane	interconnection
domain	load	stringer	manual
model	lesson	flap	wire
aircraft	learn	slat	drive
surface	result	panel	system

set of documents have been developed in the past few years, e.g. term frequency-inverse document frequency weight (if-idf) (Salton and Bucklay, 1988), Latent Semantic Indexing (Deerwester et al, 1990) or probabilistic Latent Semantic Analysis (Hoffmann, 1999). Several trials in that context and in addition diverse publications (Wei and Croft, 2006) have shown that the Latent Dirichlet Allocation (LDA) is feasible for Information Retrieval tasks. The LDA uses a probabilistic topic model to identify topics and associate words within a set of documents to a topic, which is linked to a discrete calculated probabilistic value.

Exemplarily for such results, table 1 shows four topics (out of 20) and for each topic the belonging seven top-ranked keywords. The topic model was build with the Latent Dirichlet Allocation method out of a set of documents referring to the aviation domain.

The input for LDA is a set of documents (corpus), whereas each document is represented by a "bag-of-words", that means the sequence of words within a document is not considered for following calculations. The output of LDA is a probabilistic assignment $P(d|t)$ to a topic t for each document d and a probabilistic assignment $P(w|t)$ for each keyword w to a topic t.

The topic labels have been assigned manually to provide in addition "human readable" headings for each topic. Obviously headings are not of importance for the LDA approach, but alleviate remarkable a communication with externals and "non-specialists". The latter is especially of importance in the context of "evaluation" and acceptance of the given results in a practical environment. As it turned out in several of our field studies, users (especially in a business environment, such as product development) show distrust to search results provided by a totally nebulous knowledge retrieval approach.

In order to ensure practicability and usability of the proposed solution, the respective documents used for this test-bed refer to "real" document repositories, which are in use as a knowledge repository for the development of Wing High Lift components.

2.1 Similarity Computation

Similarity, in sense of our approach, occurs by the ability to compare the probabilistic assignment to each topic between different documents. The Kullback Leibner (KL) divergence is recommended by (Steyvers and Griffiths, 2007) as an appropriate function to calculate the similarity, whereas p and q is the topic distribution to a given topic t out of the set of topics T:

$$KL(p,q) = \sum_{t \in T} P(p|t) \cdot \log_2 \frac{P(p|t)}{P(q|t)} \quad (1)$$

With respect to an appropriate post-processing of the computed results, in some cases (e.g. to generate a symmetric distance matrix for a two-dimensional result map) a symmetric measurement KL_{sym} necessary:

$$KL_{sym}(p,q) = \frac{1}{2}\left(KL(p,q) + KL(q,p)\right) \quad (2)$$

Such a similarity computation can be done among different documents in a given repository or even between a text paragraph (instead of a complete document) and documents in a given repository. The latter can be compared to a search query, e.g. to provide similar documents to a selected text passage of a document in progress or of interest.

Table 2: Top five results to a search query. Further a text summarization for each resulting document.

Document	Relevance	1.word	2.word	3.word	4.word	5.word	6.word	7.word	8.word	9.word
Movable_flap ....pdf	91 %	flap	fairing	Aug	investigation	fair	ll	a340	modificati on	ata
Fairing_Intefe rence...pdf	89%	flap	fairing	Aug	investigation	fair	ll	a340	modificati on	ata
Flap_Track_F airing...pdf	85%	flap	fairing	Aug	investigation	fair	ll	a340	modificati on	ata
Flap_Track_ Actuator...pdf	84%	flap	fairing	Aug	investigation	fair	ll	a340	ata	track
Movable_FT F...pdf	83%	flap	fairing	Aug	investigation	fair	ll	a340	ata	movable

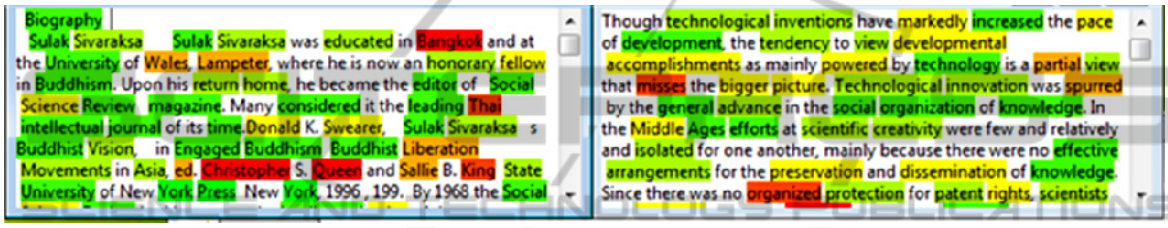

Figure 2: An abstract of the annotated preview of "Sulak Sivaraksa" (84% relevance; left) and "Social Development Theory" (91% relevance; right). Relevance was computed with KL divergence and visual assignment with function (4) (k=0.05).

The calculated divergences can then be used to rank the documents according to their similarities in e.g. a table form or a two-dimensional map (Kleiza et al., 2010).

Next to the above mentioned function the Euclidean Distance or Cosine Similarity are also possible distance functions for computing the semantic similarity by use of a topic model.

2.2 Text Summarization

As mentioned above the authors claim to have identified a need for an enhancement of the described similarity computation approach by addressing the traceablity of the KL based search results (in terms of *why is the specific document highly ranked* or *which document parts are upmost interesting)*.

In order to provide an overview for the user to indicate, why those results are similar to each other or to a given search query, several approaches have been analysed. So far, all approaches base either on a sentence (Wang et al., 2008) (Gong and Liu, 2001) or on word summarization (Dredze et al., 2008) (Liu et al., 2009) of the documents to give an insight meaning of the similarity. In most cases a word summarization is preferable, because of a shorter overview and faster cognitive understanding and filtering during a search process. Here the set of words which is considered by the summarization always comes out of the current document. Remaining words of the whole dictionary, which could be also meaningful for the summarization, are so far not taken into consideration. (Dredze et al., 2008) uses the following function to rank the words within a document, whereas c is the word (candidate) for the relevance computation and d the related document or text paragraph:

$$P(c \mid d) = \prod_{w \in d} \sum_{t \in T} P(c \mid t) P(w \mid t) P(t \mid d) \qquad (3)$$

For instance the ten words with the highest probability could be visualized in a tag cloud or as an addition to the search result list (table 2) to give a better understanding, why the results are similar to a given query (or a text paragraph respectively). The search query input below comes up with a result list mainly with documents of the Lessons Learned section in the wing high lift device domain:

"Several cases of A340 movable flap track fairing n°5 with one case on A330 have been reported with inner skin cracked. Cracks lengths ranged from 180 to 2600 mm and are not related to accidental damage."

The relevance values are computed with the KL divergence function followed by a normalization and conversion to percentage values. Table 2 shows the top five documents ranked with the highest relevance and a word summarization computed with function (3). The probabilistic evaluation $P(c|d)$ of the words in Table 2 varies between a range of 10^{-76} and 10^{-94}. Numerous multiplications of the terms between 0 and 1 are responsible for such low values.

3 APPROACH

In many cases text summarization as shown in table 2 brings not the required usability to filter a sub-set of relevant documents fast and easy according to their importance for a given context. The equality of text summarization is often attended by the similarity computation of documents. The characteristics are often not given by a text summarization (as shown in table 2). Hence our approach bases on the idea to mark up those words directly in the resulting document, which are semantic similar. This way a visual rendering of the document with the highlighted text passages within a preview pane can be archived. We assume, that a user gets a better understanding of the deep insight of the similarity between document and search query with a highlighted preview. The semantic highlighting goes further than a syntactic highlighting of words which annotates matching search keywords in the document, like known by common search engines. Hence such an approach for semantic highlighting on probabilistic topic model needs a elaborated theoretical background.

3.1 Word Similarity

The underlying algorithm for word similarity has to be equal for all documents in order to get comparable previews of different documents. Therefore highly ranked documents (in term of the KL Divergence computation) should offer many words with high similarity and lower ranked documents should contain fewer words with high similarity and a lower similarity on average respectively.

$$D(w|d,q) = -k.\log_2 \sum_{t \in T} P(w|t)P(t|d)P(t|q) \quad (4)$$

The importance / similarity of a word w in the document d for the current query q is given by function (4). The scalar k gives the ability to stretch the values in order to refine the visualization for the user.

3.2 Visual Assignment

Even though it is not in focus of the core research of the semantic similarity approach, an appropriate visualization of results is to be expected to be a key issue for user acceptance in a practical environment. Therefore a visual assignment of an arbitrary word by its similarity calculated by function (4) to a search query has been defined.

Figure 1: Visual assignment for highlighting identified words.

Figure 1 shows an assignment of highlighted keywords to a color spectrum in the style of an "signal light". A green annotation indicates a strong, yellow a middle and red low significance. The visual assignment fades out to indicate no significance arranged behind the colors. Alternative possible annotation types could be e.g. underlining or surrounding the words, but from initial user feedback based on paper mockups it turned out, that a simple color annotation (as also shown in fig. 2) is easy to grasp and comes along with an outstanding effect. The visual assignment can be justified by modifying the scalar k to get a proper visualization. However, a fine-tuning of k in consultation with respective end users is foreseen to be part of the planned field study.

4 RESULTS

During a test the number of topics was set to 100 and the topics have been computed out of a set of about 1000 Wikipedia documents by usage of LDA algorithms. Stopwords have been removed and a lemmatization of the words has been done. The identification of significant words was computed by usage of function (4) and the following query input:

"People need a broad range of skills in order to contribute to a modern economy and take their place in the technological society of the twenty-first century. An ASTD study showed that through technology, the workplace is changing, and so are the skills that employees must have to be able to change with it."

Exemplarily for the computational output two excerpts are shown in Figure 2. Both arbitrary selected previews provide identified and highlighted words. The left preview (Sulak Sivaraksa) highlights in intensive green words in relation to three topics: religion ("Buddhism"), social ("University", "Social", "Interactual") and press ("Magazine", "Journal", "Press"). On the other side the right preview (Social Development Theory) highlights intensive green words in relation to Technology ("Technology", "Knowledge", "Development", "Effective") and Social ("Social", "Organization") topics. As a conclusion it can be stated that different words out of different topics have been identified, which match the meaning of the given query-input very well in any topic (social, technologic) or have generally a significant meaning for the specific document (e.g. press and religion).

It has to be stated that the enhancement of a semantic search engine by implementation of a word similarity algorithm can be definitely used to identify semantic relevant sections and text passages within a document. Even though the current implementation is still at the prototype stage, benefits to be gained by its usage already become clearly visible, especially with respect to the implemented visual assignment.

5 OUTLOOK

As one of the next major milestones an evaluation of the approach in a field study is planned to validate the concept on operational level. This field study will be conducted by potential key users belonging to a product development department in the aviation sector. By picking up the group of engineers, designers and physicians for aerodynamics, working in the area of Aircraft Design as target users on the one hand and their document repositories as a test-bed on the other, the authors want to ensure practicability and usability of the proposed solution.

On a technical level several improvements are foreseen either: Next to LDA, other topic models will be tested and compared by each other in order to receive an overview how well the proposed approach works upon them. For instance the authors expect to have even better results by usage of the HMM-LDA (Griffiths et al., 2005) approach instead of LDA, because the HMM-LDA not only considers the unstructured "bag-of-words" input, but also reflects the sequence of words within the respective document. By using HMM-LDA the authors expect to come up with a solution, where the intensive highlighted words will be even more concentrated on a specific text passage. Next to those examinations further optimization and refinements of the core algorithms for computation of the word similarity are foreseen in order to continuously improve the output quality.

ACKNOWLEDGEMENTS

This work has been done under the LuFo IV 2nd call -HIGHER-TE- WP4200 research programme funded by the Airbus S.A.S. We wish to acknowledge our gratitude and appreciation to all the project partners for their contribution during the development of various ideas and concepts presented in this paper.

REFERENCES

Bahrs, J. et al., 2007. Wissensmanagement in der Praxis - Ergebnisse einer empirischen Untersuchung: Empirische Studien in der Wirtschaftsinformatik 1. ed., Gito.

Blei, D.M., Ng, A.Y. & Jordan, M.I., 2003. Latent dirichlet allocation. The Journal of Machine Learning Research, 3, 993-1022.

Deerwester, S. et al., 1990. Indexing by latent semantic analysis. Journal of the American Sociaty for informations science, 41(6), 391-407.

Dredze, M. et al., 2008. Generating summary keywords for emails using topics. In Proceedings of the 13th international conference on Intelligent user interfaces. Gran Canaria, Spain, pp. 199-206.

Gong, Y. & Liu, X., 2001. Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the *24th annual international ACM SIGIR conference* on Research and development in information retrieval. New Orleans, United States, pp. 19-25.

Griffiths, T.L. et al., 2005. Integrating topics and syntax. Advances in neural information Processing Systems, 17, 537-544.

Hofmann, T., 1999. Probabilistic Latent Semantic Analysis. Proceedings of uncertainty in artificial intelligence, 289-296.

Kleiza, K. et al., 2010. Integrated Semantic Search in the Product Development Phase. to be published in *Proceedings of the 16th International Conference* on Concurrent Enterprising.

Liu, S. et al., 2009. Interactive, topic-based visual text summarization and analysis. In Proceeding of the 18th ACM conference on Information and knowledge management. Hong Kong, China, pp. 543-552.

Salton, G. & Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513-523.

Steyvers, M. & Griffiths, T., 2007. Probabilistic topic models. Handbook of Latent Semantic Analysis, 424–440.

Wang, D. et al., 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. Singapore, pp. 307-314.

Wei, X. & Croft, W.B., 2006. LDA-based document models for ad-hoc retrieval. In Proceedings of the 29th annual international *ACM SIGIR conference on Research and development in information retrieval*. Seattle, USA, pp. 178-185.