Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 551 – 559

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Supporting contextualized information finding with automatic excerpt categorization

Ricardo Kawase[a,*], Patrick Siehndel[a], Bernardo Pereira Nunes[b]

[a]*L3S Research Center, Leibniz University Hannover, Appelstr. 9a, 30167 Hannover, Germany*
[b]*Department of Informatics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro/RJ – Brazil, CEP 22451-900*

Abstract

The volume of information on the Web is constantly growing. Consequently, finding specific pieces of information becomes a harder task. Wikipedia, the largest online reference Website is beginning to witness this phenomenon. Learners often turn to Wikipedia in order to learn facts regarding different subjects. However, as time passes, Wikipedia articles get larger and specific information gets more difficult to be located. In this work, we propose an automatic annotation method that is able to precisely assign categories to any textual resource. Our approach relies on semantic enhanced annotations and the categorization schema of Wikipedia. The results of a user study show that our proposed method provides solid results for classifying text and provides a useful support for locating information. As implication, our research will help future learners to easily identify desired learning topics of interest in large textual resources.

Peer-review under responsibility of KES International.

Keywords: Annotation; Categorization; Wikipedia; Learning Support;

1. Introduction

Since the rise of the Web 2.0, the volume of information available has significantly grown. Users have become the core contributors to the Web information space, producing a wide range of content and transforming it into the main source of information to the most variety of topics.

In fact, the advent of the Web 2.0 has also created a cultural change in how people interact, communicate and acquire knowledge.

* Corresponding author. Tel.: +49-(0)511-762-19715 ; fax: +49-(0)511-762-19712.
E-mail address: kawase@L3S.de

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.136

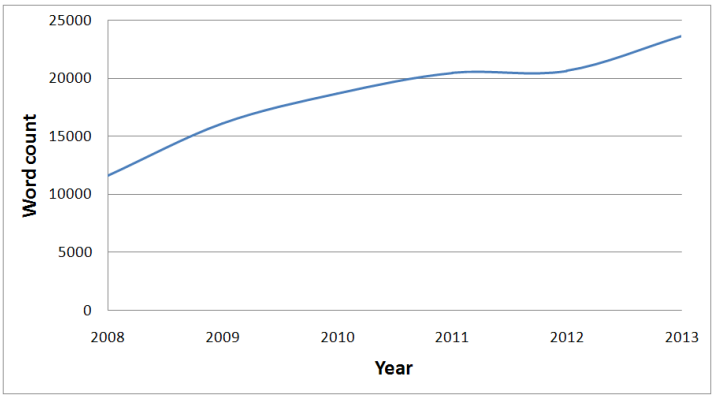

Fig. 1. Word count of Wikipedia artilcle 'Barack Obama' in the last five years.

A recent report from Pew Research Center[1] shows evidences of such user behaviors, where 92% of the adult users utilise the Web to perform online search and exchange e-mails. Although the increasing amount of the information arguably creates a richer Web, it also brings drawbacks and challenges. As more information is available, the more difficult it becomes to find, select and consume relevant contents.

This is particularly a problem in students' learning process, where a flood of information might hinder their understanding. For instance, students with attention deficit disorder may suffer even more, since they have difficulties in sustaining attention, fail to give attention to details and are easily distracted. In this manner, if the provided content is focused solely on the students' interests, or if students can focus only on excerpts of texts that they are interested in, then, the chances to get distracted is decreased.

We can illustrate the increase of information through a look at Wikipedia[2], a free encyclopedia created collaboratively by people who use it. Currently, Wikipedia has almost 30,000 active contributors, 4,4 million articles and registers over 3 million edits per month[3]. If we take into consideration the Wikipedia article of Barack Obama[4], we will find that, in terms of length (word count), his Wikipedia article page has duplicated in the last 5 years (See Figure 1) from 11,609 words (September 12, 2008) to 23,653 (September 12, 2013).

Over time, as in any other Web page, Barack Obama's article will definitely change and new content will be added. The constant growth of information may hinder the consume of information by its users and, therefore, new forms to access it must be provided. Thus, if a user is interested solely on Obama's association to *Sport* or *Education*, instead of pointing to his Wikipedia page, we must point the user to the excerpt of text in his Wikipedia page related to those topics of interest. Since an article in Wikipedia serves as a starting point for learning, delimiting its topics would facilitate and improve learning experience.

In this light, our main motivation in this paper is to extract topic-relevant information from Web pages and provide to end users an overview of the contents based on the topics they address. Our research is closely related to text segmentation, summarization and classification, however, differently from previous works in the field, our method relies on entity extraction and semantic classification.

Concisely, given a textual resource and a topic of interest, our method describes the input by selecting only the topic-relevant information. To achieve this goal, we identify the main topic subject for each paragraph.

Our high-level topic classification relies on the Wikipedia top categories which contain a broad coverage of topics that are maintained by the overall agreement of millions of contributors. This topic classification provides readers a sense making categorization that is digestible and manageable. While other approaches like clustering and Latent

[1] `http://www.pewinternet.org/Static-Pages/Trend-Data-(Adults)/Online-Activites-Total.aspx` accessed on Sept. 12
[2] `http://www.wikipedia.org`
[3] `http://stats.wikimedia.org/EN/SummaryEN.htm`
[4] `http://en.wikipedia.org/wiki/Barack_Obama`

Dirichlet allocation (LDA)[1] provide means for categorization and recommendation of items, they do not support the end user in understanding the topics.

The remainder of the paper proceeds as follows. Section 2 reviews closely related literature to our approach. Section 3 describes in details our step-by-step approach to classify text segments. Section 4 provides a user study to validate the usefulness of our approach and Section 5 presents the results. Finally, Section 6 discusses the outcomes of our approach and points out future directions.

2. Related work

In the last years, as the amount of information in the Web grew, new methods to classify and find information emerged. In this light, a lot of research has been done to improve the task of automatically classifying documents.

TF-IDF weighting is arguably the most accepted approach to begin with text classification[2,3,4,5]. This well known strategy turns documents into a list of weighted terms that facilitates the representation of the documents. It relies on the assumption that the most representative terms of a document occur many times in the document's text and, at the same time, occur only in a small set of the available documents.

A great part of the literature on text classification is based on machine learning approaches and rely on dimensionality reduction[6] or on probabilistic topic models[7]. These strategies begin with training set of manually (positive examples of classification) annotated documents. Based on these sets, algorithms find existing patterns in given documents. These patterns are later on used to automatically identify classes[8,2,9].

In all these works, document classification has been proven to be an important component that supports information retrieval tasks. In fact, document classification is key to ensuring quality of any digital library. In previous works, we have presented novel approaches for document classification[10] as well as competence classification[11,12], and the importance of these features in learning scenarios.

However, to the best of our knowledge, there is not much research done in the direction of classification of text segments. Classifying text segments (in our case paragraphs) brings the same benefits of text annotations. Text and digital annotations are well known to be a great learning companion[13]. In fact, active reading[14] is arguably the most common learning process, where learners read, and at the same time perform other activities such as writing and annotating the text. These annotations serves different purposes[15] nevertheless, all of them support the readers to refind information[16].

In this work, we try to solve a not yet explored problem, namely the annotation cold-start scenario. As we have discussed in the Section 1, online documents get larger and specific pieces of knowledge get harder to be discovered. Annotations have the potential to support readers in directly finding the desired topics. However, previous approaches discussed in this section fail to properly identify a set of terms that might effectively help readers. Paragraphs are rather short and do not contain enough information for machine learning approaches.

Our proposed method has several advantages over these approaches. First, we enrich the pieces of text with semantic information. Second, we use a common world-wide adopted knowledge base (Wikipedia). Finally, our method can be applied to any type of text, from any repository and is not hindered by cold-start scenarios.

3. Approach

In this section, we present a step-by-step approach for classifying text segments in Web documents. The approach is divided into: (i) annotation; (ii) categorization; and (iii) aggregation (See Figure 2).

Briefly, the first step is responsible for an entity identification and extraction process that links entities found in a Web document (e.g. Wikipedia articles) to relevant Wikipedia references. The second step is key in our approach, being responsible for traversing knowledge bases and identification of possible text segment categories. Finally, the last step generates an overall score to the categories found in the second step to create a final text segment profile.

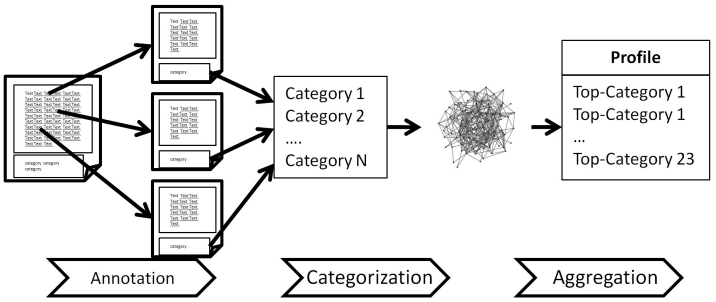

Fig. 2. Classification workflow.

3.1. Annotation

The first step in our approach consists of identifying entities (i.e. creating links) in articles to relevant Wikipedia references. Although Wikipedia articles are strongly interlinked, usually one hyperlink does not reoccur within the same page.

To perform the first task, we use the WikipediaMiner[17] tool, a Web annotation service that is responsible for identifying all mentions of entities that can be linked to Wikipedia articles. Basically, the WikipediaMiner algorithm consists of two phases. First, it detects and disambiguates words in the text that represent links to Wikipedia. To disambiguate, WikipediaMiner relies on machine learning algorithms that take into consideration the context of the word.

Next, based on the first phase, the algorithm creates links from the disambiguated words to Wikipedia articles. Only those words that are considered to be relevant for the whole document are linked to the corresponding articles. The goal of the whole process is to annotate a given document in the same way as a human would link a Wikipedia article. Early publications of Wikipedia Miner reports recall and precision of almost 75%, whether the system is evaluated on Wikipedia articles or "real world" documents[17].

3.2. Categorization

In this step, we extract Wikipedia categories of each entity that has been identified in the previous step. Note that Wikipedia has 25 main categories that comprehensively cover all existing knowledge fields (all Wikipedia categories are connected in a directed graph where 25 of them represent top categories). For each Wikipedia category, we follow the path to parent categories, up to the root category. In some cases, this procedure results in the assignment of multiple top level categories to a single entity. Following the parent categories (which are closer the root category), we compute values of distance and siblings categories, resulting in each entity receiving 25 categories' scores. In fact, there are different approaches that can be applied to walk Wikipedia's category graph. To achieve best results and accurately assign weights to each of the 25 categories, we experimented with different graph walk and weighting strategies. A detailed evaluation is provided in Section 3.3.

3.3. Aggregation

Finally, in the *aggregation* step, we perform a linear aggregation over all of the scores for a given paragraph in order to generate the final profile.

We used the Wikipedia category graph for relating one paragraph to the 25 main Wikipedia categories. The dataset used contains 593,125 different categories. Each of these categories is linked to one or more of the main categories. Table 1 shows some statistics of the used graph.

We used two different graph walking algorithms for computing the relation of a category to the main categories. Both strategies follow a top-down approach that pre-computes main category weights for each article. The main difference between the two approaches is the size of the generated tree for each main category. The relation of an article to the main categories is based on a depth first walk through the Wikipedia category graph: the algorithm

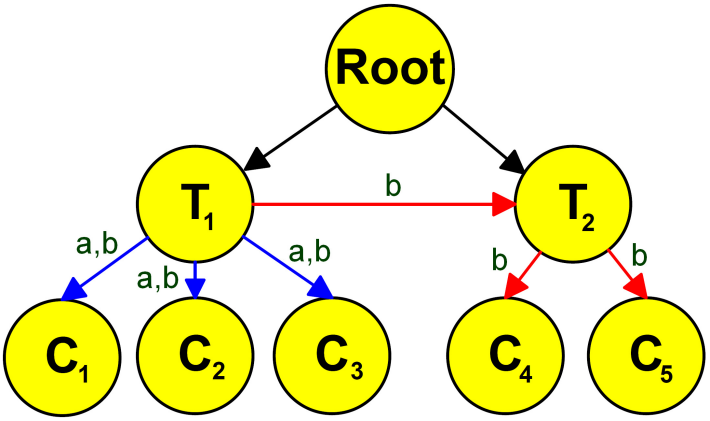

Fig. 3. Subcategories of TC_1 for strategies 'A' and 'B'.

Table 1. Statistics on the Wikipedia Category Graph.

# of Categories	593,125
# of Category-Subcategory links	1,306,838
avg. # of Subcategories	2.2
# of Page-Category Links	11,220,967
avg. # of Pages per Category	18.9

remembers the distance from the root node, and follows only sub-category links of which the distance is larger (*strategy A*) or equal (*strategy B*) to the current distance to the root node.

Figure 3 shows a small graph that consists of a root, two top level categories (T_i) and 5 normal categories (C_i). When strategy 'A' is applied on this graph, category T_1 will contain all articles that are related to the categories C_1, C_2, C_3 and T_1. The category T_2 will not be part of C_1 because there exists another way with equal length from the root to T_2. When strategy 'B' is used on this graph, all categories will be seen as part of T_1. We empirically evaluated both strategies on random articles and found out that strategy 'B' provided better results (detailed report of these experiments is out of the scope of this paper). Thus, we used this strategy for the remaining experiments in this paper.

By following only links that match this pattern, we make sure not to include the entire category graph (and all articles) for each main category. Additionally, we avoid loops by storing visited nodes and not visiting these nodes again. For the subcategories that are reachable through the category graph, we get the corresponding articles that belong to the categories. With this approach, we get a relation map in which every category is related to many articles and, in which most articles are related to many categories.

A basic profile for a given paragraph consists of weights for all of the main categories. The final weight θ of a topic $t \in T$ (top 25 Wikipedia categories) for a paragraph $o \in O$ is given by Equation 1:

$$\theta(o_i, t_k) = \sum_{e \in o_i}^{l=|e|} \left(\sum_{c_j(e_l)}^{j=|c_j(e_l)|} w(c_j, t_k) \right), \tag{1}$$

where e are the entities annotated in a given paragraph o, $c(e)$ are the Wikipedia categories for $e \in o_i$ and w is a weight given to the link between a category c_j and a top-category t_k. To complete, we define the weight used in our experiments (see Equation 2).

There are big variances between the different categories. Categories like 'Mathematics', 'Agriculture' or 'Chronology' are relatively weakly represented. This leads to a classification in which these categories are underrepresented as well.

To achieve a more precise classification, we calculate the weight of the top categories taking into account the relative probability of an article belonging to one of the main categories. Additionally, we assume that a longer distance to one of the main categories can be interpreted as a weaker relation to that category. The calculation is shown as Equation 2.

$$w(t_k, c_j) = \frac{1}{P(t_k)} * \frac{1}{\delta(t_k, c_j)} \tag{2}$$

where $P(t_k)$ indicates the popularity of a given top-category and δ is the distance of a category c_j to the top-category t_k. To measure the performance in this experiment, we calculate the average rank of the correct main category inside the profile vector.

In the end, each given paragraph receives a profile that consists of a 25-sized weighted vector, representing how relevant the paragraph is to each of the Wikipedia categories. Based on this profile, it is possible to identify to which extent a pararaph approaches each specific topic of interest.

4. User study

The goal of the user study is to validate the accuracy, and consequently, the usefulness of the profiling method. In this section, we describe the experimental setup to evaluate the proposed method.

4.1. Dataset

In order to validate the outcomes of our method, we setup a user study with a few selected articles from Wikipedia. We considered a scenario where learners would look for information regarding countries and politicians. To this end, we accessed the Wikipedia articles contained in two Wikipedia lists, namely the lists of countries[5] and the list of current members of the United States House of Representatives[6].

These two lists combined link to 728 Wikipedia articles containing information about the aforementioned lists. In most cases, these articles are very extensive containing plenty of information covering many, and sometimes all different topics (defined in Wikipedia). On top of this sample, we applied our profiling methods for each paragraph in these articles.

In total, we extracted and annotated 34,095 paragraphs. In average, the paragraphs have 500.62 characters, 74.08 words and 4.86 sentences. In these paragraphs, the method performed 588,204 annotations, linking them to 64,524 unique Wikipedia articles. The distribution of number of entities found per article is depicted in Figure 4. To illustrate it, the top three articles with most number of entities found are respectively the articles of the countries Ukraine[7], Portugal[8] and France[9]. The logarithmic distribution of paragraphs per topics is illustrated in Figure 5. As expected, following the nature of the chosen articles (Politicians and Countries), the top identified topics were *Politics*, *Culture*, *Geography* and *History*.

4.2. Participants and setup

To evaluate our method, we set up our evaluation on CrowdFlower[10], a crowdsourcing platform. With CrowdFlower we are able to reach a broader unbiased audience to judge our outcomes.

For the evaluation, we selected from each category 50 random paragraphs (note that not all categories were assigned to this amount of paragraphs). In total, the selected sample to be evaluated consisted of 869 paragraphs.

[5] http://en.wikipedia.org/wiki/List_of_sovereign_states
[6] http://en.wikipedia.org/wiki/List_of_current_members_of_the_United_States_House_of_Representatives_by_seniority
[7] http://en.wikipedia.org/wiki/Ukraine
[8] http://en.wikipedia.org/wiki/Portugal
[9] http://en.wikipedia.org/wiki/France
[10] http://crowdflower.com/

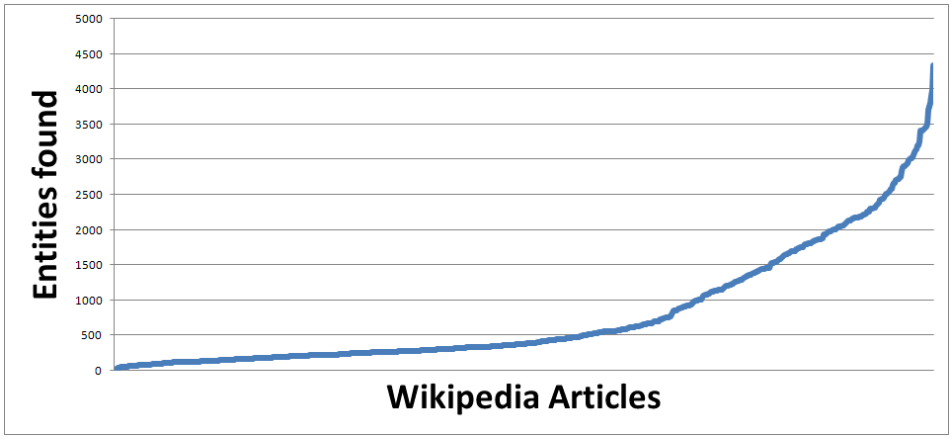

Fig. 4. Distribution of annotated entities across the articles in the experiments' sample.

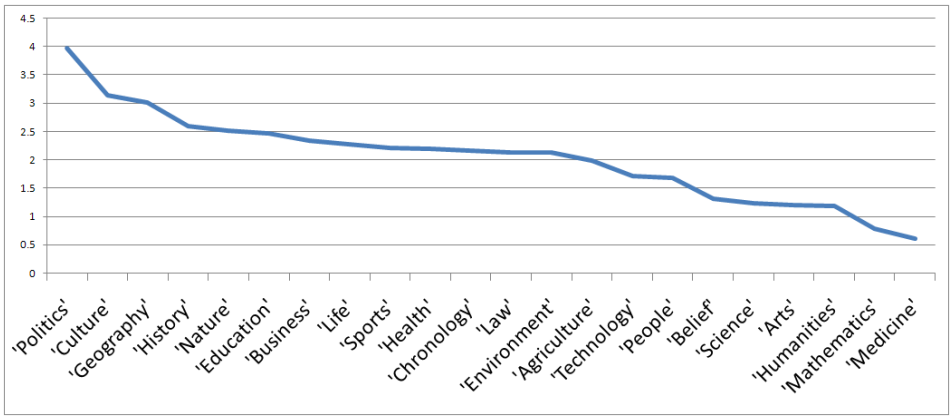

Fig. 5. Logarithmic distribution of paragraphs across the categories.

The evaluation process consisted of a questionnaire in a 5-point Likert scale model where participants were asked to rate the agreement of the suggested categories to a given paragraph according to relevance. For each paragraph, we presented the three top ranked categories found by our proposed methods. The judgments collected provide us a quality review of the proposed annotation and categorization strategies.

Participants were first instructed to read the paragraph, and were also aware of the existing ranking of the three suggested categories (being the first one the most relevant suggestion). The participation in the evaluation was limited to English native speakers and each participant was asked to evaluate a maximum of 50 paragraphs. Each participant was presented with one paragraph at a time, randomly drawn.

5. Results

In total, we recruited 53 participants and each item was evaluated by at least three different participants. The participants preferences are depicted in Figure 6. Only one category suggestion was voted with 1 (not related). The great majority of votes report that participants found the suggested category *relevant* to *very relevant*. In fact, these results sum up to over 95% of all votes.

A closer look into the wrongly classified paragraph and its identified entities indicates that the main cause of the misassignment was the ambiguity of the terms. This can be worked around by tuning the annotation process with more restrictive parameters that improves the contextualization.

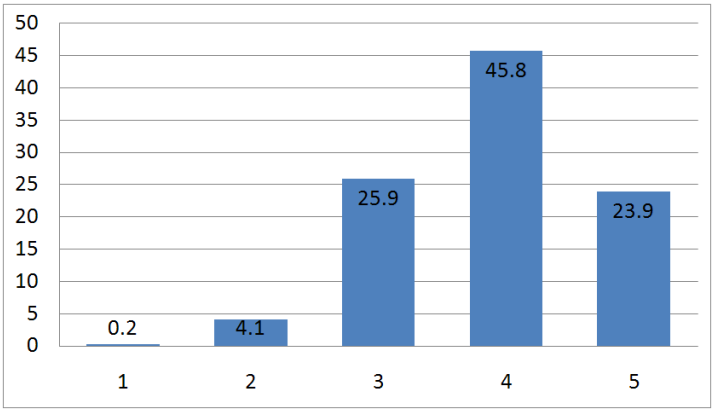

Fig. 6. Evaluation results: percentage of votes regarding the relevance of suggested categories for a given paragraph in a 5-point Likert scale.

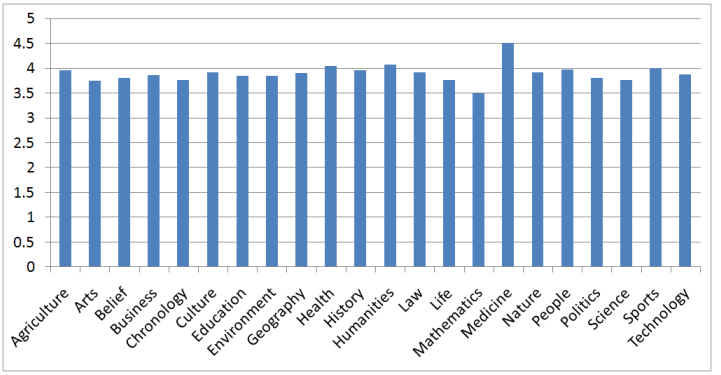

Fig. 7. Evaluation results: Average rating per category.

In order to observe if there is any significant difference between categories, we grouped the results as depicted in Figure 7. The results show no significance between any of the categories. This means that the proposed method works equally for any different content disregarding its topic.

6. Conclusions

In this paper, we proposed a method for automatically annotating excerpts of text. Our approach relies on semantic enhanced annotations and Wikipedia's categorization schema - arguably the most complete knowledge base currently available online. The text segments' categorization supports readers in quickly accessing desired information. Here, we presented the first evaluation in order to access the quality of the categorization. The results show that the vast majority of categories assigned to paragraphs were correctly related. These results are very promising and demonstrate the applicability of our methods.

Regarding the potential improvement in learning scenarios, in previous works, we have demonstrated that contextualized clues improve information finding[16] on the Web and in online courses[13]. In fact, contextual clues are appreciated by learners that consume online textual resources[13]. Selecting the relevant excerpts of text from extensive textual resources holds the same allegory of in-context clues; it points the reader directly to relevant content in the text. In this light, we hypothesize the applicability of your proposed method in real learning scenarios will greatly assist learners, paving the path for future work.

Our planned future work is divided in two directions. First, we will validate the effectiveness of the categorization in supporting learners to find information in real case scenarios. Second, we plan to upgrade the classification method

in order to annotate paragraphs with a different granularity other than Wikipedia top categories. A preview of our method is available online[11].

Acknowledgements

This work has been partially supported by the European Commission under QualiMaster (ICT 619525).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.. Latent dirichlet allocation. *J Mach Learn Res* 2003;**3**:993–1022. URL: http://dl.acm.org/citation.cfm?id=944919.944937.
2. Sebastiani, F.. Machine learning in automated text categorization. *ACM Comput Surv* 2002;**34**:1–47. URL: http://doi.acm.org/10.1145/505282.505283. doi:10.1145/505282.505283.
3. Joachims, T.. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. MA, USA: Kluwer Academic Publishers; 2002. ISBN 079237679X. URL: http://portal.acm.org/citation.cfm?id=572351.
4. Kolcz, A., tau Yih, W.. Raising the baseline for high-precision text classifiers. In: Berkhin, P., Caruana, R., Wu, X., editors. *KDD*. ACM. ISBN 978-1-59593-609-7; 2007, p. 400–409. URL: http://dblp.uni-trier.de/db/conf/kdd/kdd2007.html#KolczY07.
5. Soucy, P., Mineau, G.W.. Beyond tfidf weighting for text categorization in the vector space model. In: Kaelbling, L.P., Saffiotti, A., editors. *IJCAI*. Professional Book Center. ISBN 0938075934; 2005, p. 1130–1135. URL: http://dblp.uni-trier.de/db/conf/ijcai/ijcai2005.html#SoucyM05.
6. Rendle, S., Schmidt-Thieme, L.. Pairwise interaction tensor factorization for personalized tag recommendation. In: *Proceedings of the third ACM international conference on Web search and data mining*; WSDM '10. New York, NY, USA: ACM. ISBN 978-1-60558-889-6; 2010, p. 81–90. URL: http://doi.acm.org/10.1145/1718487.1718498. doi:10.1145/1718487.1718498.
7. Diaz-Aviles, E., Fisichella, M., Kawase, R., Nejdl, W., Stewart, A.. Unsupervised auto-tagging for learning object enrichment. In: *EC-TEL*; vol. 6964 of *Lecture Notes in Computer Science*. Springer. ISBN 978-3-642-23984-7; 2011, p. 83–96. URL: http://dblp.uni-trier.de/db/conf/ectel/ectel2011.html#Diaz-AvilesFKNS11.
8. Joachims, T.. Text categorization with support vector machines: Learning with many relevant features. 1998. URL: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.6124.
9. Moschitti, A., Basili, R.. Complex linguistic features for text classification: A comprehensive study. In: McDonald, S., Tait, J., editors. *ECIR*; vol. 2997 of *Lecture Notes in Computer Science*. Springer. ISBN 3-540-21382-1; 2004, p. 181–196. URL: http://dblp.uni-trier.de/db/conf/ecir/ecir2004.html#MoschittiB04.
10. Kawase, R., Fisichella, M., Nunes, B.P., Ha, K.H., Bick, M.. Automatic classification of documents in cold-start scenarios. In: *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*; WIMS '13. New York, NY, USA: ACM. ISBN 978-1-4503-1850-1; 2013, p. 19:1–19:10. URL: http://doi.acm.org/10.1145/2479787.2479789. doi:10.1145/2479787.2479789.
11. Kawase, R., Siehndel, P., Nunes, B.P., Fisichella, M.. Automatic competence leveling of learning objects. In: *ICALT 2013: 13th IEEE International Conference on Advanced Learning Technologies (ICALT), Beijing, China*. 2013, .
12. Kawase, R., Siehndel, P., Nunes, B.P., Fisichella, M., Nejdl, W.. Towards automatic competence assignment of learning objects. In: Ravenscroft, A., Lindstaedt, S.N., Kloos, C.D., Leo, D.H., editors. *EC-TEL*; vol. 7563 of *Lecture Notes in Computer Science*. Springer. ISBN 978-3-642-33262-3; 2012, p. 401–406. URL: http://dblp.uni-trier.de/db/conf/ectel/ectel2012.html#KawaseSNFN12.
13. Pereira Nunes, B., Kawase, R., Dietze, S., Bernardino de Campos, G., Nejdl, W.. Annotation tool for enhancing e-learning courses. In: Popescu, E., Li, Q., Klamma, R., Leung, H., Specht, M., editors. *Advances in Web-Based Learning - ICWL 2012*; vol. 7558 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. ISBN 978-3-642-33641-6; 2012, p. 51–60. URL: http://dx.doi.org/10.1007/978-3-642-33642-3_6. doi:10.1007/978-3-642-33642-3_6.
14. Adler, M.J., van Doren, C.. How to read a book. Simon and Schuster; 1972, .
15. Kawase, R., Herder, E., Nejdl, W.. A comparison of paper-based and online annotations in the workplace. In: *EC-TEL*. 2009, p. 240–253.
16. Kawase, R., Papadakis, G., Herder, E., Nejdl, W.. The impact of bookmarks and annotations on refinding information. In: *HT*. 2010, p. 29–34.
17. Milne, D., Witten, I.H.. Learning to link with wikipedia. In: *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*. New York, NY, USA: ACM. ISBN 978-1-59593-991-3; 2008, p. 509–518. URL: http://dx.doi.org/10.1145/1458082.1458150. doi:10.1145/1458082.1458150.

[11] http://twikime.l3s.uni-hannover.de/all/twikime_twikify.php