To the Point: A Shortcut to Essential Learning

Ricardo Kawase, Patrick Siehndel
Leibniz University of Hanover & L3S Research Center
Appelstrasse 9, 30167 Hannover, Germany
Email: {kawase, siehndel}@L3S.de

Bernardo Pereira Nunes
Department of Informatics - PUC-Rio
Rio de Janeiro, RJ - Brazil
Email: bnunes@inf.puc-rio.br

***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 Wikipedia's categorization schema. The results of a user study shows 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.**

***Keywords*-Topic coverage, Topic extraction, Discussion forum, Topic recommendation, Forum assessment**

I. Introduction

Since the rise of the Web 2.0, the volume of information available has significantly grown. Users have become the core contributor 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.

This is particularly a problem in the student's 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, fails to give attention to details and are easily distracted. In this manner, if the provided content is focused on the students interest or if the student can only focus in excerpts of texts that s/he is interested in, then the chances to get distracted and sustain attention is decreased.

This is particularly a problem in the student's 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, fails to give attention to details and are easily distracted. In this manner, if the provided content is focused on the students interest or if the student can only focus in excerpts of texts that s/he is interested in, then the chances to get distracted and sustain attention is decreased.

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 [1]. In previous works, we have presented novel approaches for document classification [2] as well as competence classification [3], [4], 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. Thus, in this paper, we present the work towards a system that can automatically classify and identify topic-relevant excerpts within large texts.

II. Approach

The approach is divided into: (i) annotation; (ii) categorization; and (iii) aggregation.

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 a overall score to the categories found in the second step to create a final text segment profile.

A. Annotation

The first step in our approach consists of discovering entities (links) in articles to relevant Wikipedia references. Although Wikipedia articles are strongly interlinked, usually one hyperlink does not reoccur within the same page. It is common practice in Wikipedia to link only the appearance of a term.

Therefore, we first fully annotate the articles to detect all mentions of entities that can be linked to other Wikipedia articles. For this purpose, we use the WikipediaMiner [5] service as an annotation tool. The WikipediaMiner approach consists of two basic steps: first, detected words are disambiguated using machine learning algorithms that take the context of the word into account.

Our Wikipedia dataset contains over 4 million articles, covering almost all knowledge domains. In order to maximize the identification of links in Wikipedia articles, we minimized the confidence parameter. With more information available, more accurate will be the aggregation step.

B. Categorization

In following step, *categorization*, we extract the categories of each link (entity) that has been identified in the previous step. For each Wikipedia category, we follow the path of all parent categories, up to the root category. In some cases, this procedure results in the assignment of several 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 different graph walk and weighting strategies.

C. 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. 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. In the end, each given paragraph receives a profile that consists of a weighted 25 sized 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.

III. User Study

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 particular topics. We annotated 728 Wikipedia articles containing information regarding politicians and countries. In total we extracted and annotated 34,095 paragraphs.

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. A random sample 869 paragraphs of was evaluated by volunteers.

A. Results

In total, we recruited 53 participants and each item was evaluated by at least three different participants. 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.

IV. 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 learners 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.

The future work is divided in two directions. First, we plan to upgrade the classification method in order to annotate paragraphs with a different granularity other than Wikipedia top categories. Second, we will validate the effectiveness of the categorization in supporting learners to find information in real case scenarios. A preview of our method is available online[1].

References

[1] E. Diaz-Aviles, M. Fisichella, R. Kawase, W. Nejdl, and A. Stewart. Unsupervised auto-tagging for learning object enrichment. In *EC-TEL*, volume 6964 of *Lecture Notes in Computer Science*, pages 83–96. Springer, 2011.

[2] R. Kawase, M. Fisichella, B. P. Nunes, K.-H. Ha, and M. Bick. Automatic classification of documents in cold-start scenarios. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, WIMS '13, pages 19:1–19:10, New York, NY, USA, 2013. ACM.

[3] R. Kawase, P. Siehndel, B. P. Nunes, and M. Fisichella. Automatic competence leveling of learning objects. In *ICALT 2013: 13th IEEE International Conference on Advanced Learning Technologies (ICALT), Beijing, China*, July 2013.

[4] R. Kawase, P. Siehndel, B. P. Nunes, M. Fisichella, and W. Nejdl. Towards automatic competence assignment of learning objects. In A. Ravenscroft, S. N. Lindstaedt, C. D. Kloos, and D. H. Leo, editors, *EC-TEL*, volume 7563 of *Lecture Notes in Computer Science*, pages 401–406. Springer, 2012.

[5] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518, New York, NY, USA, 2008. ACM.

[6] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 81–90, New York, NY, USA, 2010. ACM.

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