TagTheWeb: Using Wikipedia Categories to Automatically Categorize Resources on the Web

Jerry Fernandes Medeiros[1(✉)], Bernardo Pereira Nunes[2], Sean Wolfgand Matsui Siqueira[1], and Luiz André Portes Paes Leme[3]

[1] Department of Applied Informatics, Federal University of the State of Rio de Janeiro (UNIRIO), Rio de Janeiro, RJ 22290-240, Brazil
{jerry.medeiros,sean}@uniriotec.br
[2] Department of Informatics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ 22451-900, Brazil
bernardo@ccead.puc-rio.br
[3] Institute of Computing, Fluminense Federal University, Niterói, RJ 24210-310, Brazil
lapaesleme@ic.uff.br

Abstract. Identifying topics associated with a set of documents is a common task for many applications and can be used to improve various tasks involving documents on the Web, such as search, retrieval, recommendation, and clustering. To address this problem, this paper introduces a tool, called TagTheWeb, as a proposition of a generic classification method, that relies on the knowledge expressed by the taxonomic structure of Wikipedia, based on the generation of a fingerprint through the semantic relation between nodes of the Wikipedia Category Graph. TagTheWeb can be used as a WEB interface or as an API to classify any text based resource.

Keywords: Text categorization · Semantic web · Wikipedia Categories · Category Graph

1 Introduction

Wikipedia is the most substantial encyclopedia freely available on the Web. It has been developed and curated by a large number of users over time and represents the common sense about facts, people and the broadest type of topics currently found on the Web.

One of the outstanding features of Wikipedia is the categorization system used to index its internal content. Very briefly, there are a finite number of top categories that represents the whole Wikipedia content. These top categories, as well as their subcategories, are not fixed and are maintained and curated by Wikipedia users.

A. Gangemi et al. (Eds.): ESWC 2018 Satellite Events, LNCS 11155, pp. 153–157, 2018.
https://doi.org/10.1007/978-3-319-98192-5_29

The primary purpose of this research is to create a general-purpose classification tool based on Wikipedia Categorization scheme that can categorize text-based content on the Web, for instance, scientific articles, web pages or even posts on social media. At the current stage, it is possible to categorize any textual content in different languages via a web interface or API.

2 Related Work

The depth and coverage of Wikipedia has attracted the attention of many researchers who have used it as a knowledge resource for several tasks, including text categorization [2], predicting document topics [8] and computing semantic relatedness [3,6,7].

Halavais and Lackaff [4] quantitatively compared the distribution of 3,000 Wikipedia articles coded into Library of Congress categories with a distribution of published books. They found substantial overlap between Wikipedia categories and topics from other encyclopedias. Kittur [5] demonstrated a simple technique for determining the distribution of topics for articles in Wikipedia, mapping all items to the top categories. The process was based on building the Category Graph of Wikipedia and counting the edges on the shortest paths from the categories of an article to the top categories of Wikipedia. Farina [1] improved this by penalizing edges followed in the wrong direction concerning the hierarchy. Strube and Ponzetto [9] developed a system named WikiRelate!. They used data from Wordnet, Wikipedia, and Google for computing degrees of semantic similarity and reported that Wikipedia outperforms Wordnet. They used different measures for computing semantic relatedness and showed good results with the one based on paths.

3 TagTheWeb - Approach Overview

Our primary goal is to take advantage of the Wikipedia body of knowledge to automatically categorize any text-based content on the Web according to the collective knowledge of Wikipedia contributors. A processing chain to generate a generic categorization was developed based on three steps: (i) Text Annotation; (ii) Categories Extraction; (iii) Fingerprint Generation.

As the basis for our approach, we consider the relationships of Wikipedia Categories as a directed graph. Let $G = (V, E)$ be a graph, where V is the set of nodes representing Wikipedia categories, and E is the set of edges representing the relationships between two categories.

To make it simple to understand, let us illustrate the steps.

3.1 Text Annotation

When dealing with the Web of Documents, we are primarily working with unstructured data, which, in turn, hinders data manipulation and the identification of atomic elements in texts. To alleviate this problem, information

extraction (IE) methods, such as Named-Entity Recognition (NER) and name resolution, are employed. These tools automatically extract structured information from unstructured data and link them to external knowledge bases in the Linked Open Data cloud (LOD), which is DBpedia in this case.

For instance, after processing the following Web resource using an IE tool: "I agree with Barack Obama that the whole episode should be investigated.", the entity "Barack Obama" is annotated, classified as "person" and linked to the DBpedia resource <http://dbpedia.org/resource/Barack_Obama>, where structured information about the entity is available.

3.2 Categories Extraction

Given the entities found in the previous step as a starting point, the categories extraction step begins by traversing the entity relationships to find a more general representation of the entity, i.e., their categories. All categories associated with the entities identified in the source of information are extracted.

For instance, for each extracted and enriched entity in a Web resource, we explore the relationships through the predicate [dcterms:subject], which by definition represents the categories of an entity. In that sense, to retrieve the topics, we use SPARQL query language for RDF over the DBpedia SPARQL, where we navigate up in the DBpedia hierarchy to retrieve broader semantic relations between the entities and its topics.

3.3 Fingerprint Generation

The goal of this step is to assign a set of main topics within Wikipedia Categories to a given web resource.

Our approach consists of navigating in the Category Graph from each category extracted in the previous step towards the top of the graph by all the shortest paths between the category and the main topics.

Each time the source category reaches one of the top-level categories, we update the influence of this top category in the composition of the resource classification.

Based on the influence of each main topic category in the resource, we generate a fingerprint, which represents the calculated categorization as a multidimensional vector, making it easy to retrieve and compare documents. For Instance, using a straightforward similarity metric such as cosine.

As a formal definition, let us denote I as the set of categories related to a web resource d, found in the category extraction step. C is the set of all Categories in Wikipedia, and M is the set of categories that represent the main topics. $G = (V, E)$, where $I \subset V; C \subset V; M \subset V$; and $M \subset C$. The parameter t is defined to indicate the broadest t levels to be considered in the set of M. If t is 1, only the main topics previously defined are considered; if t is 2, any category 1 edge away in the graph is also considered as one of the main topics, as represented in Algorithm 1. An example of the tool can be seen in Fig. 1.

Algorithm 1. Fingerprint Generation

```
procedure GenerateFingerprint(G, M, I, t, w)
    E ← a map from a list of categories m ∈ M
    for i ∈ I do
        S ← the set of shortest paths between i and any category in M
        for s ∈ S do
            B ← the set of last t vertices in path s
            for b ∈ B do
                E[b] ← E[b] + w
    return E
```

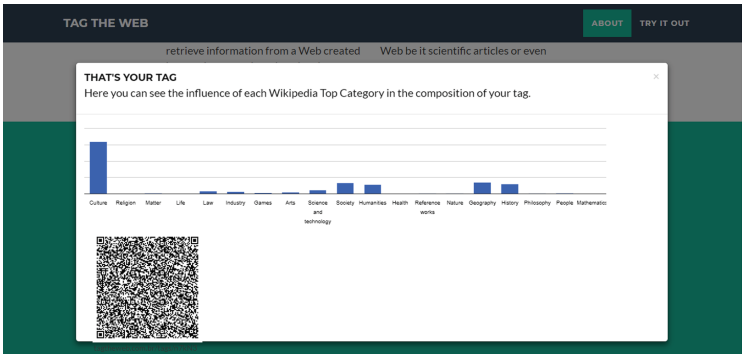

Fig. 1. Categorization for a given page of Obama's Twitter

4 Preliminary Evaluation and Results

The first validation of this work was the analysis of the fingerprint in posts of question and answers communities. Stack Exchange is a network of 133 Q&A (Question and answers) communities on topics in varied fields, each community covering a specific theme, where questions, answers, and users are subject to a reputation award process.

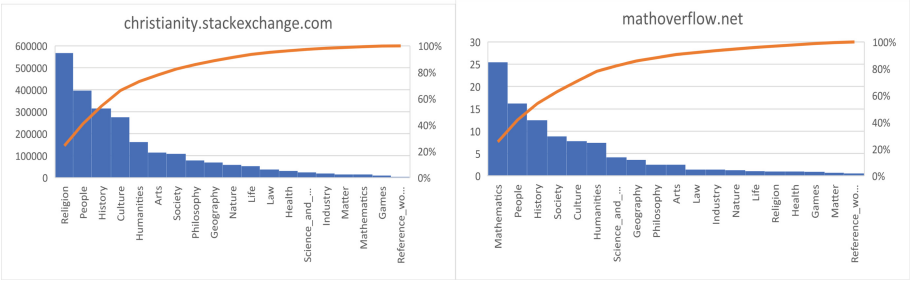

Fig. 2. Distribution of topics along the questions on stack exchange communities

We relied on an anonymized dump of all user-contributed content on the Stack Exchange network, extracted on August 31st. We selected four representative communities on stack exchange to perform this evaluation: (1) Biology; (2) Christianity; (3) Law; and (4) Math. For each row in the Post.xml file of each one of these communities, we executed the three steps of the chain described in Sect. 3. The topic distribution for each community is displayed in Fig. 2.

5 Conclusion and Future Works

This paper introduced TagTheWeb, a tool to automatically categorize resources on the web based on the Wikipedia Category Graph. A preliminary empirical evaluation shows promising results as we can reliably classify question and answers on communities that cover specific themes. As future works, we intend to test TagTheWeb in other scenarios to fine-tune the algorithm. We are also conducting experiments with humans to identify whether they agree or not with the categorization generated by the tool. TagTheWeb is publicly available at tagtheweb.com.br, and the API documentation can be found at https://documenter.getpostman.com/view/1071275/tagtheweb/77bC7Kn.

References

1. Farina, J., Tasso, R., Laniado, D.: Automatically assigning Wikipedia articles to macrocategories (2011)
2. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge. In: AAAI, vol. 6, pp. 1301–1306 (2006)
3. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: IJCAI, vol. 7, pp. 1606–1611 (2007)
4. Halavais, A., Lackaff, D.: An analysis of topical coverage of Wikipedia. J. Comput.-Mediat. Commun. **13**(2), 429–440 (2008)
5. Kittur, A., Chi, E.H., Suh, B.: What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1509–1512. ACM (2009)
6. Milne, D.: Computing semantic relatedness using Wikipedia link structure. In: Proceedings of the New Zealand Computer Science Research Student Conference. Citeseer (2007)
7. Ponzetto, S.P., Strube, M.: Exploiting semantic role labeling, Wordnet and Wikipedia for coreference resolution. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp. 192–199. Association for Computational Linguistics (2006)
8. Schönhofen, P.: Identifying document topics using the Wikipedia category network. Web Intell. Agent Syst.: Int. J. **7**(2), 195–207 (2009)
9. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using Wikipedia. In: AAAI, vol. 6, pp. 1419–1424 (2006)