Expert Systems With Applications 61 (2016) 327–342

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Identification of category associations using a multilabel classifier

Julian Szymański*, Jacek Rzeniewicz

Faculty of Electronics, Telecommunications and Informatics, Department of Computer Architecture, Gdańsk University of Technology, Poland

ARTICLE INFO

Article history:
Received 11 January 2016
Revised 24 May 2016
Accepted 27 May 2016
Available online 1 June 2016

Keywords:
Wikipedia
Associations mining
Semantic networks
Categorisation
SVM

ABSTRACT

Description of the data using categories allows one to describe it on a higher abstraction level. In this way, we can operate on aggregated groups of the information, allowing one to see relationships that do not appear explicit when we analyze the individual objects separately. In this paper we present automatic identification of the associations between categories used for organization of the textual data. As experimental data we used a network of English Wikipedia articles and their associated categories, that have been preprocessed by a dedicated filtering method for noise reduction. The main contribution of the paper is the introduction of the method based on supervised machine learning for mining relations between these categories. We describe existing in the literature category proximity metrics as well as introduce three new ones, based on observing the properties of a multilabel Support Vector Machine classifier. The first metric uses classifier predictions, the second uses its errors, and the third is based on its model. Comparison to the existing state-of-the-art methods, and to manual assessments, confirm that the proposed methods are useful and are more flexible than typical approaches. We show how different metrics allow us to introduce new significant relations between categories. Aggregated results of mining categories' associations have been used to build a semantic network that shows a practical application of the research. The proposed method for finding associations can be extended with using other approaches than SVM classification, and can find (other than presented in the paper) applications for mining categories in text repositories. Eg.: it can be used for extending the prediction of the rating in recommender systems or as a method of missing data imputation.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Wikipedia is currently the largest encyclopedia in the world. Its English version alone contains over fifty times more entries than Encyclopedia Britannica, the second largest English language encyclopedia. As of March 2014, English Wikipedia contains over 4.6 million pages[1] and continues to expand, without losing its impact.

Apart from being a rich source of information readable for humans, Wikipedia is a tremendous data repository that is the interest of many researches related to knowledge engineering. One can download it as a whole – complete dumps[2] are published regularly – and subject it to any experiments imaginable. In the last years a number of interesting research projects have been conducted with Wikipedia, for example the SGI Project[3] where links, events and dates were analyzed to visualize the world's history in the last two centuries, or the one using Wikipedia to compute semantic relatedness between words or texts (Gabrilovich & Markovitch, 2007).

Automatic processing of Wikipedia articles often requires a category similarity metric (Chen, Ma, & Zhang, 2009). For example, this was the case in the approach to topic classification described in (Farina, 2010), as well as in the semantic relatedness algorithm WikiRelate! (Strube & Ponzetto, 2006). In both approaches, similarity between categories was approximated by the distance in a category graph. However, this approach does not yield very good results. First of all, it lacks stratification – the number of category pairs assigned the highest similarity (lowest distance) equals the number of edges in the graph. Secondly, such a metric yields a fair fraction of false positives – vaguely connected categories are marked as similar. This is caused by the fact that links between categories, already marked in Wikipedia are not strictly taxonomic, and often short paths are observed for utterly unrelated classes.

Other applications that would be possible with a good similarity metric pertain to Wikipedia analysis and visualization. For example, clustering performed with a similarity measure that goes beyond a category graph could certainly bring some interesting

* Corresponding author.
E-mail addresses: julian.szymanski@eti.pg.gda.pl (J. Szymański), jacek.rzeniewicz@eti.pg.gda.pl (J. Rzeniewicz).

[1] http://stats.wikimedia.org/EN/TablesWikipediaEN.htm
[2] http://dumps.wikimedia.org/enwiki/
[3] http://www.sgi.com/go/wikipedia/

http://dx.doi.org/10.1016/j.eswa.2016.05.039
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

results. Visualisation of Wikipedia edits, performed with a state-of-the-art similarity metric algorithm (Holloway, Bozicevic, & Börner, 2007), confirms this point.

More aspects of similarity metric application are related to Wikipedia's users and editors. As for users – browsing the encyclopedia could become more engaging when the relevant category links to content that is thematically related. In the case of editors, the way they manage the category system could be improved. Currently it is the editors' job to create category nodes, and links between them. Because of its large scale, the category graph is incomprehensible for one person, and it takes time until a new category is linked with similar ones. With a good similarity metric, an editor could receive recommendations about classes that might be related with the currently edited page or category.

The main goal of the research presented in this paper is to develop approaches to computing the similarity between categories that organize textual repository. A total of five metrics were implemented, evaluated and compared. Two of them were based on ideas published in other works, described in Section 2. The first one is based on assignments of articles to categories; in the second one, similarity is computed for classes represented by words contained in related articles. These metrics form a baseline for comparison of metrics introduced by us that have been based on observing the properties of a multilabel support vector machine (SVM) (Hearst, Dumais, Osman, Platt, & Scholkopf, 1998) classifier: first based on its predictions, second on errors, and third on its model. Section 5 describes results of the experiments conducted in order to evaluate and compare these metrics. Evaluation was performed in two ways. First, original hierarchic relationships defined in Wikipedia were used as a reference set. For each of the methods it was measured how well it restores those relations. Second, manual assessment of some of the results yielded by the metrics was performed.

The secondary contribution of research presented in this paper is the demonstration of an approach showing how Wikipedia can be automatically transformed into a semantic network. For this, relationships between categories obtained using four selected approaches to compute categories' similarity were combined. Concepts and relations of the network were exported and used in the visualisation presented in the last section.

2. Related work

2.1. Association rule mining

The problem of finding relations between elements of large datasets has been studied since the early nineties. One of the major catalysts of this area of data mining was barcode technology popularization, which allowed big retail stores to automatically collect past transactions data (basket details). (Agrawal, Imieliński, & Swami, 1993) designed an efficient algorithm for mining association rules between sets of items. For example, given sales data obtained from a large retailing company, the following rule was found: `Children's Hardlines` ↦ `Infants and Children's wear` which might be an indicator for shop managers that products belonging to these two categories should be placed close to each other. In the above case, the rule contained only one item in both antecedent and consequent. However, the algorithm also guarantees to find rules containing larger itemsets, e.g.: `Cheese, Ground beef, Ketchup` ↦ `Burger buns`.

Even though the worst case complexity is 2^m, where m denotes number of all items, in practice the algorithm runs much faster because basket sizes are most usually limited by some small number. It is also worth noting that the association rule mining algorithm produces a directed graph, so $A \to B$ does not necessarily mean $B \to A$. A number of methods have been proposed since (Agrawal et al., 1993) that improved its efficiency (Agrawal & Srikant, 1994; Brin, Motwani, Ullman, & Tsur, 1997).

With the transition of sales to the Internet, baskets are often replaced by session logs: one itemset would contain products related to pages a user visited within a single session. What is more, online stores allow for returning customer identification (e.g. using cookies). Thus not only more data is available to process, but also a new field of recommender systems emerged. Association rules mining was adapted in the 1:1Pro system to construct personal profiles based on past activity of individual customers (Adomavicius & Tuzhilin, 2001). Also a method described in (Lin, Ruiz, & Alvarez, 2000) makes use of association rules in order to come up with personalized user recommendation.

2.2. Mining direct relationships between Wikipedia categories

(Holloway et al., 2007) proposed an approach to computing relationships between Wikipedia categories grounded on an idea similar to association rules mining. In this solution articles correspond to baskets, and categories assigned to an article form a single itemset. However there are two simplifications in the proposed approach. First, relationships are calculated only between individual categories. Second, the resulting graph is undirected. Similarity between categories C_i, C_j is specified using cosine Formula 1.

$$\cos_{i,j} = \cos_{j,i} = \frac{\sum_{k=1}^{n} A_k C_i \cdot A_k C_j}{\sqrt{\sum_{k=1}^{n} A_k C_i \cdot \sum_{k=1}^{n} A_k C_j}} \tag{1}$$

where $A_k C_i$ equals 1 when article A_k is assigned to C_i and 0 otherwise. The Formula 1 can be geometrically interpreted as a cosine between vectors in an n-dimensional space, with each dimension representing one Wikipedia article.

An interesting property of Wikipedia categories is due to the imposed naming conventions[4]. Category names must follow a strict structure defined for various types of categories, like topic categories (*Law, Civilization*) or set categories (*Writers, Villages in Poland*). (Nastase & Strube, 2013) inferred new relationships between categories from their names. For example, category named *Mixed martial arts television programs* is a direct concatenation of other two categories *Mixed martial arts* and *Television programs*, therefore there is a premise that perhaps these categories are related.

A somewhat different approach to computing a similarity metric for Wikipedia categories is to infer it from similarities between articles. This solution can yield various results depending on the assumed articles similarity function. A widely used technique allowing one to compute similarity between documents, is by representing them in a Vector Space Model (Salton, Wong, & Yang, 1975) using words as features. A metric based on such representation of articles was implemented in (Szymański, Deptuła, & Krawczyk, 2013). Concretely, in the approach described in this work, TF-IDF weighting (Salton & McGill, 1986) was applied to the features and similarity between two articles A_i, A_j was computed according to Formula 2:

$$s(A_i, A_j) = \frac{\|a_i \odot b_{ij}\|_1 + \|a_j \odot b_{ij}\|_1}{\|a_i\|_1 + \|a_j\|_1} \tag{2}$$

where a_i and a_j are vectors representing articles A_i and A_j, respectively, and b_{ij} is a vector defined as follows:

$$b_{ij}[f] = \begin{cases} 1 & \text{when } a_i[f] \neq 0 \ \wedge \ a_j[f] \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

[4] http://en.wikipedia.org/wiki/Wikipedia:Category_names

Download English Version:

https://daneshyari.com/en/article/6855656

Download Persian Version:

https://daneshyari.com/article/6855656

Daneshyari.com