LINKED DATA AND PAGERANK BASED CLASSIFICATION

Michal Nykl, Karel Ježek
Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Pilsen
Univerzitni 22, 306 14 Pilsen, Czech Republic

Martin Dostal, Dalibor Fiala
NTIS, Faculty of Applied Sciences, University of West Bohemia, Pilsen
Univerzitni 8, 306 14 Pilsen, Czech Republic

This is preprint version, original version is:

NYKL, Michal, JEŽEK, Karel, DOSTAL, Martin and FIALA, Dalibor. Linked Data and PageRank based classification. In: *IADIS International Conference Theory and Practice in Modern Computing 2013 (part of MCCSIS 2013)*. Praha: IADIS Press, 2013, pp. 61-64. ISBN: 978-972-8939-94-6.

ABSTRACT

In this article, we would like to present new approach to classification with Linked Data and PageRank. Our research is focused on classification methods that are enhanced by semantic information. The semantic information can be obtained from ontology or from Linked Data. DBpedia was used as source of Linked Data in our case. Feature selection method is semantically based so features can be recognized by nonprofessional users because they are in a human readable and understandable form. PageRank is used during feature selection and generation phase for expansion of basic features into more general representatives. It means that feature selection and processing is based on a network relations obtained from Linked Data. The features can be used by standard classification algorithms. We will present the promising preliminary results that show the easy applicability of this approach to different datasets.

KEYWORDS

Linked Data, PageRank, classification, feature selection

1. INTRODUCTION

Document classification is an important part of document management systems and other text processing services. Today`s methods are usually statistically oriented that means a big amount of data is required for training phase of these classification algorithms. The preparation of sufficient classification training sets and proper feature selection methods is challenging task even for domain specialists. So the common solution of this problem is based on relatively comprehensive corpuses that contain a lot of documents divided into different classification classes. Statistical methods are trying to discover relations between terms and classification classes during training phase.

Our approach recognizes interesting keywords and expands them using semantic information obtained from Linked Data. For example based on a feature *MySql* we can do the feature expansion into *databases* without explicit occurrence of this word in document content. The classification phase can process these parent concepts and use them for correct pairing of documents and classification classes.

Next, we will explain basic principles of Linked Data and introduce the primary variant of PageRank that will be used in next sections. Related work is discussed in Section 2 and our approach to feature selection is described in Section 3. Preliminary evaluation of our method was performed with 20 News Groups dataset. The results are presented in Section 4.

1.1 Linked Data and PageRank

The concept of Linked Data was first introduced by Tim Berners-Lee (Berners-Lee, 2006). He formulated four rules for machine readable content on the Web:

- Use URIs as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information using the standards (RDF*, SPARQL).
- Include links to other URIs so that they can discover more things.

More specific is the idea of Linked Open Data, which is based on the presumption of freely published data without restrictions in usage or additional fees.

The PageRank algorithm was developed in 1998 by Page and Brine (Brine, 1998) as approach to web pages ranking through exploring hyperlinks structure on the Internet. The importance of each web page depends on the number and PageRank value of all web pages that link to it. This approach also known as "Random surfer walking" has been studied and improved for citation analysis (Ma, 2008). Our modified version of PageRank (1) corresponds to its matrix definition (Langville, 2006) where $P_x(a)$ is a value of node a in iteration x, d is a damping factor usually set to 0.85, V is a set of all nodes in the graph, U is a set of nodes with link to node a, D is a set of all dangling nodes and w_{ij} is weight of link from node i to node j.

$$P_{x+1}(a) = \frac{(1-d)}{|V|} + d * \left(\left(\sum_{u \in U} \frac{P_x(u) * w_{ua}}{\sum_{v \in V} w_{uv}}\right) + \frac{\sum_{s \in D} P_x(s)}{|V|}\right) \quad (1)$$

2. PREVIOUS WORK

Document classification can be defined as content-based assignment of one or more predefined categories (classification classes) to documents. We can distinguish two phases in document classification processing, the learning phase and the classification phase. In the learning phase user define categories by giving training documents for each of these categories. Quality improves with increasing number of training documents. This is a weak point of document classification because the solid training collection is required.

There have been many supervised learning techniques for document classification. Some of these techniques include Naive Bayes, k-nearest neighbor, vector approaches e.g. Rochio, support vector machines, boosting (Schapire, 1999), rule learning algorithms (Cohen, 1996), Maximum Entropy and Latent semantic analysis.

DBPedia was used as a source of Linked Data presented in this article. We use a local copy of Linked Data stored in our relation database for performance purposes, but SPARQL endpoint could also be used. DBPedia is semantically enriched Wikipedia that was successfully employed previously for computing semantic relatedness of documents. WikiRelate! (Strube, 2006) combines path based measures, information content based measures, and text overlap based measures. Explicit Semantic Analysis (Gabrilovich, 2007) uses machine learning techniques to explicitly represent the meaning of a text as a weighted vector of Wikipedia-based concepts.

Another approach to document classification (Wang, 2005) proposed term graph model as improved version of the vector space model. The aim of this model is to represent the content of a document with relationship between keywords. This model enables to define similarity functions and PageRank-style algorithm. The vectors of PageRank score values were created for each document. The rank correlation and the term distance were used as a similarity measures to assign document to classification class. An alternative approach to document classification uses hypernyms and other directly related concepts (Bloehdorn, 2004; Ramakrishnanan, 2003). Next step in document classification can be marked as feature expansion with additional semantic information from ontology (De Melo, 2007). This approach (De Melo, 2007) is exploiting the external knowledge for mapping terms to regions of concepts. For exploration of related concepts, the traversal graph algorithm is used.

3. FEATURE SELECTION

Feature selection is the most important part of our approach. This method and its connection with Linked Data and PageRank consist of following steps:

1. Basic features are selected from documents on the base of the TFIDF (other methods, e.g. X^2, can be used too).
2. The features of each document are mapped to Linked Data nodes (identified with URI). The mapping is based on the full or partial compliance between feature (term) and name of the node in a corresponding language. This enables to form the first version of the graph.
3. One step expansion of the graph by Linked Data is executed. The Graph expansion illustrates Fig. 1, where the original (basic) nodes are marked as *I* and the expanded nodes are marked as *II*. The weights of all links are assigned. Variants of weighting see below (Fig. 2).
4. PageRank algorithm is applied to this graph.
5. If at least one of the last added nodes have a higher PageRank score, than any of the older nodes we continue with step 3, otherwise the nodes with the highest scores represent the selected features.

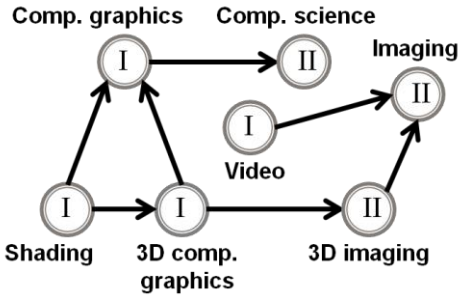

Figure 1. Graph expansion with PageRank

We have investigated three options for initialization of the PageRank algorithm (see Fig. 2), where the steps of node expansion in a graph are marked with *I*, *II* and *III*:

a) All edges are assigned the same weight equal to 1.
b) The basic nodes are advantaged with self-citation edge with increased weight.
c) The basic nodes are advantaged with self-citation edge. The edges to the new nodes are penalized base on the quadratic distance of path from the basic node.

Our evaluation of this three possibilities shows, that the variant *c)* achieves the best results due to effective limitation in expansion of the basic nodes. The other versions requires explicit limiting criterion for graph expansion. In the next section, this variant was used for evaluation of our feature selection approach.

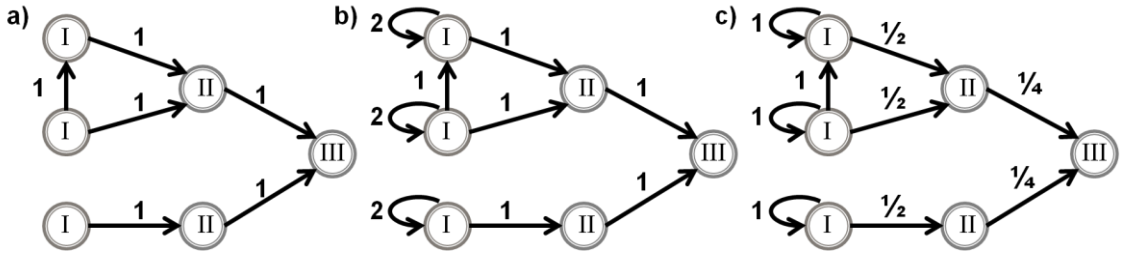

Figure 2. Initialization of the PageRank

4. RESULTS

For evaluation purposes the subset of 20 News groups collection was used. The reason was that our source of Linked Data was not sufficient to distinguish between similar categories in 20 News groups collection like *comp.os.ms-windows.misc* and *comp.windows*. Another problem was overtraining that occurs with approximately 100 training documents for 1 category.

Our approach will be completely based on the path length in a graph of nodes from Linked Data. But for preliminary results and evaluation purposes, we decided to compare our feature selection method and standard statistical approach with the same vector space classification algorithm (Rocchio). The comparison (see Fig. 3) is done with macro-averaging $F_1(\beta=1)$ measure. The number of testing documents is determined as 20% of training documents.

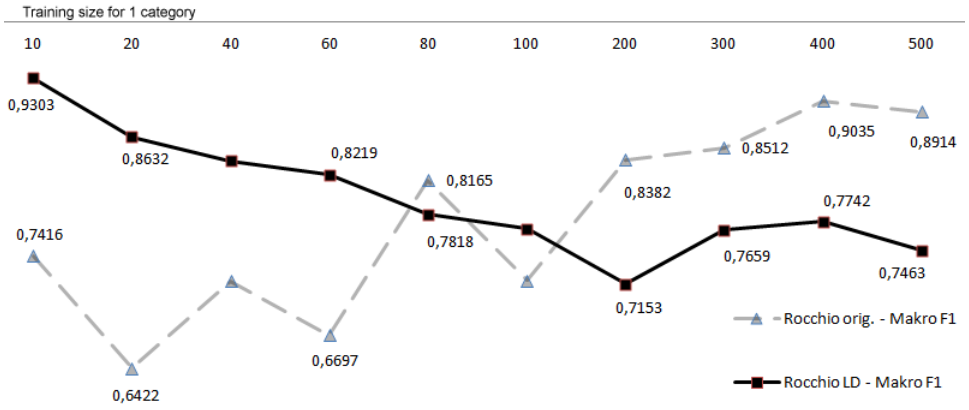

Figure 3. Graph – F1 measure for Rocchio classification algorithm

5. CONCLUSION

Our method for document classification with Linked Data is promising especially in tasks with inadequate training sets or for quick filtering of existing documents. In those cases the training phase could be very expensive and inappropriate waste of time for user. Our method allows the definition of assigning categories using only a single node from Linked Data with automatic expansion on both sites – category definition and feature selection. In the future, we would like to eliminate the overtraining problem and we would like to create a solid method for document classification directly based on the graph analysis.

This work was supported by the grants GAČR P103/11/1489 and NTIS CZ.1.05/1.1.00/02.0090.

REFERENCES

Berners-Lee, T., 2006. URL: http://www.w3.org/DesignIssues/LinkedData.html, date: 2006-07-27, cited: 2013/01/12.

Bloedhorn, S. and Hotho, A., 2004. Boosting for Text Classification with Semantic Features. *WebKDD'04*. Seattle, USA.

Brine, S. and Page, L., 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, Vol. 30, Issues 1-7, pp 107-117.

Cohen, W. and Singer Y., 1996. Context-sensitive learning methods for text categorization. *ACM SIGIR* '96.

De Melo, G. and Siersdorfer, S., 2007. Multilingual text classification using ontologies. *29th EC on IR*, Rome, Italy.

Gabrilovich, E. et al, 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis.

Langville, A.N. et al., 2006. *Google's PageRank and Beyond: The Science of Search Engine Ranking*. USA.

Ma, N. et al, 2008. Bringing PageRank to the citation analysis. *Information Processing & Management*, 44, 800–810.

Schapire R. and Singer Y., 1999. BoosTexter: A boosting-based system for text categorization. *Machine Learning*.

Strube, M. and Ponzetto, S.P., 2006. WikiRelate! Computing semantic relatedness using Wikipedia. *AAAI'06*, USA.

Wang, W. et al, 2005. Term Graph Model for Text Classification. *ADMA'05*. Wuhan, China, pp. 19-30.