ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Using Wikipedia with associative networks for document classification

N. Bloom[1,2], M. Theune[2] and F.M.G. De Jong[2]

1- Perrit B.V., Hengelo - The Netherlands

2- University of Twente, Enschede - The Netherlands

Abstract. We demonstrate a new technique for building associative networks based on Wikipedia, comparing them to WordNet-based associative networks that we used previously, finding the Wikipedia-based networks to perform better at document classification. Additionally, we compare the performance of associative networks to various other text classification techniques using the Reuters-21578 dataset, establishing that associative networks can achieve comparable results.

1 Introduction

An associative networks is a connectionist model based on work in cognitive psychology [1, 2, 3]. It models the connections between observations. Each observation has its own node in the network, which is in turn linked to other observations. When using words in a text as observations, associate networks can be used to find links between texts. This approach has in earlier research proven quite successful in document categorization [4]. Due to their low computational complexity and low information requirement, associatieve networks can be used in large text categorization and classification problems in live environments where documents are prone to be added, removed or changed. For example, associative networks can be used to provide a hierarchical structure to browse through documents in a large corporate library.

In this paper we introduce a new method of constructing associative networks which yields improved results over our previous method in terms of quality. We tested the performance of our new method on the Reuters dataset and compared the results to those of alternative classifiers.

In Section 2, we describe constructing an associative network and in Section 3 we discuss related work. Next in Section 4 we describe our experiment and compare the results to those of other algorithms that were tested on the same dataset. We present our conclusions and ideas for future work in Section 5.

2 A Wikipedia-based associative network

Associative networks are modelled as graphs, with each node representing an observation. In our WordNet-based model [4], we used synsets[1] as observations, while in our Wikipedia-based model we use Wikipedia articles. Edges in the network represent conceptual connections between those observations. The weight

[1] A set of synonyms describing a single concept.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

of those edges represents how closely connected two concepts are. Because most concepts are only related to a small number of other concepts, the associative network is a sparse graph.

We can use an associative network to analyse a text and determine key concepts. Each word in the text is linked to one or more concepts, which are activated based on the frequency of the word in the text. This activation is then spread along the edges to connected nodes, with more activation being spread to more closely connected nodes. These nodes in turn can spread activation and so on. Different methods can be used to spread the activation [4] but regardless of the method used, an activation pattern is created: a list of the activation that each node has received after the activation has finished spreading. In this pattern, concepts which are strongly linked to multiple words that are frequently used in a text will receive a high activation. By comparing activation patterns, we are able to tell which articles are closely related and which are not.

In our earlier work we constructed our Associative Networks based on Princeton WordNet [5, 6]. WordNet was chosen because it provides a network of concepts linked by conceptual meaning, which is what an associative network uses to find relationships between words through the comparison of activation patterns. Some of those relationships may be stronger than others, but WordNet does not provide this information, leaving us with no option but to intialize all connections with the same strength. For this reason, training of the network is required to settle the weights to more accurately represent the actual relationship between concepts. If we can improve the quality of the initial weights however, the performance with the same amount of training should improve.

To improve the initial configuration of the network, in this paper we look into using Wikipedia as a source for the construction of the associative network.

2.1 Network construction

Wikipedia consists of connected articles, each explaining a concept and providing links to related concepts. Thus, like WordNet, it could be used as a basis for associative networks. At first glance, Wikipedia appears to offer less information than WordNet, not having a type (such as hypernym, synonym or antonym) associated with each relationship. However, with Wikipedia articles we can gather additional information as to how strongly the two concepts are linked, based on how often the linking term is used in the article.

We cannot determine the strength of the connection between linked articles purely by the number of times one article links to another, because Wikipedia authors are encouraged not to create multiple copies of a link to the same concept within an article to improve readability. Furthermore, though this goes against Wikipedia's manual of style, some links may not be relevant to the topic of the article, but may simply be an unrelated word that happens to be used somewhere in the article.

A better indication of how strongly two articles are linked, is the number of times the title of the target article is used, even when it is not a link. However, this disregards synonyms and descriptions that also refer to the target article.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Since this use of synonyms and descriptive terms is exactly the way associative networks compare articles, it makes sense to use an associative network to judge how closely two linked Wikipedia articles are related. But this leaves us with a boot-strapping problem - we need an associative network to establish the connection between two articles, and to create that associative network, we need an associative network.

To resolve this conundrum, we create our associative network in two steps. First, we create and train a basic associative network using Princeton WordNet, as we did in our earlier work [4]. This associative network relies purely on the WordNet links and training. We then use this network to create an associative network based on Wikipedia.

Potentially, this process could be carried out again, building a new associative network using the improved Wikipedia-based associative network, but this repetition of moves was considered to be outside of the scope of this research – our goal is to improve associative networks. If a single stepped solution does not create a good associative network, there is no reason to assume a second step using an equal or inferior associative network would offer further improvements.

2.2 Classification and training

Associative networks can be used for document classification in the following way. First, we create an activation pattern for every document in the training set. For each category we take the activation pattern of all articles in that document class, averaging over them. This gives us an activation pattern for the entire category – as terms common to the category will be more frequently activated than terms not in the category, this provides an accurate overview of the concepts that are relevant to each category.

In our previous work, the associative network would classify or categorize documents into a single category, the best match out of all available categories within a hierarchy of categories [4]. In the Reuters set, documents can belong to more than one category, which means some adjustment of our algorithm is required. To deal with multiple categorization, we created a merged category activation pattern for each category, consisting of the combined and averaged activation patterns of each article in that category. This means an article that is present in two categories is used in the category activation pattern of both.

To allow the associative network to sort documents into multiple categories, a category qualification score was also added to each category. The category qualification score is a threshold value established during training. An article is matched to a category if the match between the article's activation pattern and the category activation pattern exceeds the category qualification score.

Training was done in the same way as our earlier work [4], though with a much larger training set. Depending on whether or not the associative network identified the correct classes, positive or negative reinforcement was applied through back-propagation to the paths between the actual text input and the matching between the document's activation pattern and the category activation

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

pattern. Thus, certain links in the associative network would be strengthened while others would be weakened.

3 Related work

The Reuters dataset has been used countless times to compare text classification methods. Rather than discussing them all in detail, we will merely give a short listing of the methods drawn from [7] which we use for comparison in Section 4. The naive Bayesian classifier [8] is a probabilistic model of the term density, commonly used for text classification and information retrieval. The Rocchio algorithm [9] is another popular learning method based on relevance feedback. We also compared to a k-nearest neighbour classifier [10] and the C4.5 decision tree / rule learner [11]. Finally we compared to the best Support Vector Machine result from [7].

Though we are the first to use Wikipedia to construct associative networks, we are not the first to use Wikipedia's interconnected set of articles as a base for other algorithms. Mihalcea [12] used Wikipedia for word sense disambiguation. Internal links in Wikipedia and the alternative text for those links were used to identify word forms. Ni et al. [13] used Wikipedia's multi-lingual links to extract relationships between terms in different languages. Other work featuring both Wikipedia as a source and the Reuters dataset as test data is [14]. Their goal is information retrieval based on queries instead of classification and their methodology as a result is different from our associative networks, but the underlying principle of using knowledge inherent in Wikipedia to get a conceptual understanding of texts from the Reuters set is the same.

4 Reuters dataset experiment

To be able to compare our algorithm against competing systems, we tested it using the Reuters-21578 dataset compiled by David Lewis. We used the "ModApte" split, a standard split of the Reuters set, which consists of a training set of 9603 documents and a test set of 3299 documents over 90 different categories. Specifically, we compared our algorithm against the various results by [7] which used the exact same dataset.

4.1 Results

Table 1 lists the micro-F1 scores[2] of the other algorithms listed in Section 3. Table 2 shows our own results on the same Reuters corpus. We generated a micro-F1 score for a WordNet-based and a Wikipedia-based associative network. The WordNet-based associative network was also used to constructing the Wikipedia-based associative network.

[2]The micro-F1 score is the average of precision and recall over all classes.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

	Bayes	Rocchio	C4.5	k-NN	SVM
Micro F1 score	72.0	79.9	79.4	82.3	86.4

Table 1: Results by others [7]

	Wordnet AN	Wikipedia AN
Micro F1 score	83.4	85.8

Table 2: Our own results

4.2 Discussion

From the results in tables 1 and 2, we can see that Associative Networks perform at a level that is on par with other text classification techniques. Though not beating support vector machines, especially the Wikipedia-based associative network managed to attain similar scores. Additionally, Wikipedia-based associative networks outperform WordNet-based associative networks by several points.

5 Conclusion and future work

We have proven associative networks are a match for other techniques, scoring similar to the top algorithms. In future research, we wish to further improve on our already promising results using Natural Language Processing [4] which allowed better disambiguation between homonyms, thereby increasing performance. This should allow associative networks to outperform all their competitors.

With Associative Networks based on Wikipedia outperforming WordNet-based versions, we furthermore prove that having greater understanding of the strength of relationships between concepts can help improve the quality of associative networks in general. Our method for qualifying the weight of article relations based on their textual content might additionally benefit from iterative improvement of the network as described in Section 2.1, which may be an interesting angle for future research as well.

The success of using Wikipedia to create connections suggests we can further improve the method by which associative networks learn. Currently, actual data is used for training, but with the quality of trained and pre-set relations being so important, different methods may be viable that lay a closer focus on teaching these relationships. We are looking into such new training methods based on the Montessori method which emphasises training on specially prepared data where only a single feature is different to learn that specific feature rather than actual data from a real environment.

Another possibility that using Wikipedia opens up is the use of multilingual data, based on how Wikipedia articles are linked to their equivalents in other languages. Similar methods without associative networks have already proven successful in this area [13] indicating that this may be a valid avenue for future

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

research. The advantage of using associative networks would be that it would not just allow articles in different languages to be classified without translation, but it would also allow the connections in each language to represent subtle differences in meaning more accurately.

All in all, associative networks have a lot of potential for further improvement, and though they currently rank on par with the best known techniques, they may surpass them in the future as there is still much room for further growth.

References

[1] G.F. Marcus. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Learning, Development, and Conceptual Change Series. Mit Press, 2003.

[2] R. C. Schank. *Dynamic Memory: A Theory of Learning in Computers and People*. Cambridge University Press, New York, 1982.

[3] R.C. Schank and R. Abelson. *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Earlbaum Assoc, 1977.

[4] N. Bloom. Using natural language processing to improve document categorization with associative networks. In *Proceedings of the 17th international conference on Applications of Natural Language Processing and Information Systems*, NLDB'12, pages 177–182, Berlin, Heidelberg, 2012. Springer-Verlag.

[5] G. A. Miller. Wordnet: a lexical database for English. *Commun. ACM*, 38(11):39–41, November 1995.

[6] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, illustrated edition edition, May 1998.

[7] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, Lecture Notes in Computer Science, pages 137–142, London, UK, 1998. Springer-Verlag.

[8] D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning*, Lecture Notes in Computer Science, pages 4–15, London, UK, 1998. Springer-Verlag.

[9] J. Rocchio. *Relevance Feedback in Information Retrieval*, pages 313–323. 1971.

[10] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1):69–90, April 1999.

[11] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[12] R. Mihalcea. Using Wikipedia for automatic word sense disambiguation. In *North American Chapter of the Association for Computational Linguistics (NAACL 2007)*, 2007.

[13] C. Ni, J.T. Sun, J. Hu, and Z. Chen. Cross lingual text classification by mining multilingual topics from wikipedia. In Irwin King, Wolfgang Nejdl, and Hang Li, editors, *WSDM*, pages 375–384. ACM, 2011.

[14] P. Malo, P.A. Siitari, and A. Sinha. Automated query learning with wikipedia and genetic programming. *Computing Research Repository*, abs/1012.0841, 2010.