International Journal of Research in Advent Technology, Vol.4, No.6, June 2016
E-ISSN: 2321-9637
Available online at www.ijrat.org

Fast Applying Threads for Short Text Classification Using Naive Bayes

Sayali S. Rasane[1] and Dipak V. Patil[2]
Department of Computer Engineering, G. E. S., R. H. Sapat College of Engineering, Management & Research, Savitribai Phule Pune University,Nashik, India.[1, 2]
E mail: rasane.sayali@gmail.com[1], dipakvpatil@.com[2]

Abstract- With the increased use of e-commerce and online communication and publishing, texts become available in a various categories like Web search snippets, forum, blog, news feeds, book and customer reviews. Therefore, successfully processing is important in many information retrieval applications. However, matching, classifying, and clustering these sorts of text data is a new challenge. Recently the focus of text analysis is shifted towards short texts. Measuring the Semantic similarity between short texts is an important task that can be used for various applications including text classification and text clustering. The main challenge in measuring the similarity lies in the sparsity. To overcome the sparsity, enriching the semantic representation of short text using external knowledge is required. As a case study tweeter dataset is used as an external knowledge for short text analysis. Here the semantic analysis measurement method is used which is intended for the real world noisy short texts. Finding related entities from a text generally consist of several sub-problems like key term extraction from texts, related entity finding for each key term and weight aggregation of related entities. Here the Naive Based theorem is used as it is effective especially when the short text is semantically noisy. When the short text contain some meaningless and misleading terms for estimating the main topic Naive Bayes works well. The proposed system is based on system proposed by Masumi Shirakawa et al. and we have used threads for implementing parallelism. Due to parallelism sped up of 89% is obtained.

Keywords: Semantic Similarity, Naive Bayes, Semantic Representation, Short Text Clustering

1. INTRODUCTION

Now days, the focus of text analysis is shifting toward short texts like micro blogs, search queries, search results, ads, and news feeds. Measuring the semantic similarity between the short texts is an important task that can be used for various applications like text clustering [1] and text classification [2]. The challenge in doing so is that the similarity between short texts lies in the sparsity. To overcome the sparsity, enriching the semantic representation of short texts using external data is required.

Wikipedia [3] or tweeter can be used as an external knowledge for short text analysis [1], [4], [5].Wikipedia is an encyclopedia having the dense link structure. Wikipedia also has the wide coverage of various entities such as named entities, domain specific entities, and emerging entities. Also the dump data of Wikipedia can be acquired from the web in free of cost. Due to these advantages, many researchers and developers are using Wikipedia for their research work. Likewise Wikipedia can be used to measure the semantic analysis. The Wikipedia based Explicit Semantic Analysis (ESA) [6] is a widely used method to measure the semantic similarity between texts of any length. This method creates a vector of related Wikipedia entities for the given text as its semantic representation and uses the vector for measuring the similarity. Hence finding related entities from a text involves some problems such as key term extraction, related entity finding for each key term, and weight aggregation of related entities. In order to solve the problem, ESA sums the weighted vectors of related entities for each word based on the majority rule. As well as Twitter is a great tool for social web mining because it is a rich source of social data due to its inherent openness for public consumption. It is a clean and well-documented API, rich developer tool, and has a broad appeal to users. Data mining in Twitter is simple and can bring significant value.

This approach is not suited for real-world noisy short texts where both the key terms and irrelevant terms occur very few times. The majority rule doesn't work well because of the insufficient information. In such a case, focusing on key terms while filtering out the noisy terms is important. Our proposed system is based on Semantic similarity measurements for noisy short texts using extended Naive Bayes by Masumi Shirakawa et al.[7]. Here we are giving the input to the dataset. The dataset we are using is a collection of tweets. The tweets are classified in different classes. The user is supposed to give input which is considered as a key term. Then we are finding out the probabilistic scores for the key terms and the related

entities by analyzing short text. After that we are measuring and analyzing the semantic similarity between the two texts using naive bayes algorithm. At the end we get the output entities with their probability scores.

In our proposed system we are using tweets as an external source. This method is more robust for noisy short texts because the weighting mechanism of this method is based on the Bayes' theorem. Also it can amplify the score of the related entity that is related to multiple terms in a text even if each of the terms alone is not characteristic.

2. LITERATURE

Classification of short text is an emerging area in research. The research done so far has been categorized in two basic techniques, first is short text analysis based method and second one is semantic similarity based measurements. The review of the research done is presented in this section.

2.1 Page Formatting

Some of the highlighted researches on the short text analysis are mentioned here. Ferragina and Scaiella [4] proposed a simple and fast method for entity disambiguation (Entity linking) for short texts using Wikipedia. They designed and implemented TAGME system that is able to efficiently and judiciously augment a plain-text with pertinent hyperlinks to Wikipedia pages. Meij et al. [8] also tackled entity disambiguation by using various features (e.g. anchor texts, links between articles) derived from Wikipedia for machine learning. They proposed a solution to the problem of determining what a microblog post is about through semantic linking. Also they proposed a novel method based on machine learning with a set of innovative features and show that it is able to achieve significant improvements over all other methods in terms of precision. Phan et al. [5] utilized hidden topics obtained from Wikipedia for learning the LDA classifier of short texts. They presented a general framework for building classifiers that deal with short and sparse text Web segments by making the most of hidden topics discovered from large scale data collections.

Hu et al. [9] exploited features from Wikipedia for clustering of short texts. Their work demonstrated that Wikipedia was effective as an external knowledge source. They proposed a method that employs a hierarchical three-level structure for solving the data sparsity problem of original short texts and reconstruct the corresponding feature space with the integration of multiple semantic knowledge bases. Song et al. [10] illustrated the availability of ESA for short text clustering i.e. measuring semantic distance between short texts using ESA. They have developed a Bayesian inference mechanism to conceptualize words and short text.

Sun et al. [2] utilized ESA to classify short texts with support vector machine (SVM), which is supervised machine learning technique. They proposed a probabilistic method of measuring semantic similarity for real-world noisy short texts like microblog posts. Banerjee et al. [1] also employed a similar approach to ESA for the purpose of clustering short texts. They have proposed a method of improving the accuracy of clustering short texts by enriching their representation with additional features from Wikipedia.

Xiang Wang et al.[11] used Wikipedia concept to represent short document text. The mapping from document text to Wikipedia concepts is conducted using inverted index which is built from Wikipedia articles of concepts. The traditional classification method SVM is used to perform text categorization on the Wikipedia concept based document representation. The results obtained shows that the proposed method gives better performance than traditional SVM and MaxEnt method that is based on BOW model.

PuWang et al. [12] have introduced a methodology to build a thesaurus from Wikipedia, and to leverage the thesaurus to facilitate text categorization. A unified framework has been designed to expand the "Bag of Words" BOW representation with semantic relations (synonymy, hyponymy, and associative relations), and demonstrate its efficiency in enhancing previous approaches for text classification. Wikipedia is used to improve text classification. The documents are enriched with related concepts and perform explicit disambiguation to determine the proper meaning of each concept expressed in documents. By doing so, background knowledge can be introduced into documents, which overcomes the limitations of the BOW approach. The results obtained show that this approach can achieve significant improvements with respect to the baseline algorithm.

Pu Wang and Carlotta Domeniconi[13] have made an attempt to overcome the shortages of the BOW approach by embedding background knowledge derived from Wikipedia into a semantic kernel, which is then used to enrich the representation of documents. This approach successfully achieves improved classification accuracy with respect to the BOW technique, and to other recently developed methods. This methodology is able to keep multi-word concepts unbroken; it captures the semantic closeness to synonyms, and performs word sense disambiguation for polysemy terms.

2.2 Semantic Similarity Based Measurements

Titles and headings should be in Times New Roman font (Bold) with the main body of the paper in Times

New Roman. The details of each of the paper components are described here, with a summary provided in table 1.

Some representative work on semantic similarity measurements using Wikipedia is described below:

WikiRelate by Strube et al. [14] applied several simple techniques that have been developed for Word-Net [15] to Wikipedia. Given two Wikipedia articles, they specifically compute the distance in the category structure or the overlap degree between texts. They demonstrated the effectiveness of Wikipedia-based methods on standard datasets for similarity measurements (MC, RG, and ordSim353) and core reference resolution tasks [16]. Milne et al. [17] proposed WLM that efficiently computes the similarity between two articles using the overlap degree of their incoming and outgoing links.

Graph-based methods [24], [18], [19] construct a graph in which nodes are Wikipedia articles and edges are links between articles. Using the graph they create a vector of entities [19] or directly and related entities [24], [18]. Ito et al. [20] proposed link co-occurrence analysis to speedily build an association thesaurus (determining the similarity between entities).

Hassan et al. [21] utilized cross-language links of Wikipedia to compute the similarity across languages. More recently, hybrid methods have shown to be more accurate [22], [23]. Yazdani et al. [22] utilized both text contents and links in articles, and Taieb et al. [23] leveraged text contents, categories, Wikipedia category graph, and redirection to achieve competitive or sometimes better results.

Menaka S and Radha N[25] proposed a method that uses text mining algorithms to extract keywords from journal papers. The keywords are extracted from documents using TF-IDF and WordNet. TF-IDF algorithm is used to select the candidate words. WordNet is a lexical database of English which is used to and similarity among the candidate words. The words which have highest similarity are taken as keywords. The WordNet dictionary is used to calculate the semantic distances between the keywords. The extracted keywords are having the highest similarity. Then documents are classified based on extracted keywords using the machine learning algorithms - Naive Bayes, Decision Tree and k-Nearest Neighbor.

Abdullah Bawakid et al. [26] present a system that performs automatic semantic based text categorization. The system reports on a simple analysis performed to evaluate the different implemented methods. The results obtained show that using WordNet based semantic approaches does yield to a better accuracy given that the right parameters (i.e. semantic similarity threshold) are selected.

3. PROBLEM STATEMENT

To design an algorithm using Naive Bayes classifier for efficient classification of noisy short texts.

4. PROPOSED SYSTEM

Our system is based on Wikipedia-Based Semantic similarity measurements for noisy short texts using extended Naive Bayes by Masumi Shirakawa et al.[7]. Here we are giving key term to the dataset. The dataset we are using is a collection of tweets. The tweets are classified in different classes. The user is supposed to give input which is considered as a key term. Then probabilistic scores for the key terms are found and the related entities by analyzing short text. Subsequently, we are measuring and analyzing the semantic similarity between the two text using naive bayes algorithm. At the end we get the output entities with their probability scores. To speed up the processing of method [7] we have applied multithreading techniques in java.

- Probabilistic scores are calculated for the key terms and the related entities are obtained by analyzing short texts. The probabilistic scores of the related entities are calculated as per the Equation (1).
- The Naive Bayes algorithm allows emphasizing the key terms while filtering out the noisy texts as well as it measures the semantic similarity between two texts. Finally, we get all the entities with their probabilistic scores for classification as an output.

$$P\ (c|T) = \frac{\prod_{k=1}^{K} (P(T_k \in A)P(c|T_k)}{P(T)} \quad (1)$$

Where,

- P(Tk Є A)= Probability that the term Tk i.e. the key term belongs to an article A i.e. Likelihood probability.
- P(C|TK)=Probability that the term Tk belongs to the class of the related entity i.e. Prior probability of the class.
- P(T)= Probability of key term in the entire dataset.
- P(C|T)=Final output probability of the related entity.

5. DATA FLOW DIAGRAM

A data flow Diagram is a graphical representation of all major steps, and how the data flow through the system. Following figure shows the detail data flow diagram of the system with the input and the output. In the data flow diagram of the proposed system input to the system is the input text given by the user. With the help of Nave Bayes algorithm we are computing the probabilistic scores between the key terms and the related entities extracted from the tweeter dataset. The probabilistic scores for key terms and related entities are found by analyzing short text in parallel fashion using threads.

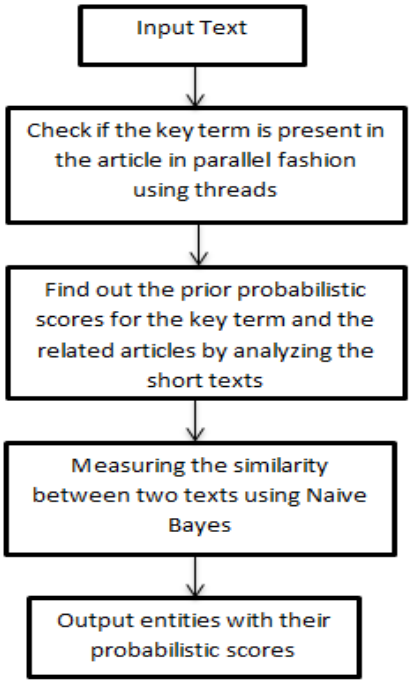

Figure 5.1: Data Flow Diagram

5.1 Algorithm and Mathematical Model

- To Extract Article :

  V ← Extract Articles(A)

  N ← Count Articles(A)

  ∀{ c ∈C}

  Do Nc ← Count Articles In Class(A, c)

  prior[c] ← Nc/N

  textc ← Concatenate Text of All article in Class(A, c)

  Do Tct ← Count Tokens of Term (textc, T)

  ∀{ t ∈V}

  Do cond prob[t][c] ← Tct + 1 ∑′ (Tct′+1)

  Return {V, prior, cond prob}

- To extract terms from random article

  W ← Extract terms from Article(V, A)

  ∀{c ∈C}

  Do score[c] ← log prior[c]

  ∀{ T ∈W}

  Do score[c] + = log condprob[t][c]

  Return argmax c ∈C score[c]

6. EXPERIMENTAL EVALUATION

We have created 5 datasets. Among these 5 datasets 3 datasets contains different number of articles about IT. These IT datasets are names as IT1, IT2 and IT3. IT1 dataset contains 3037 articles. IT2 dataset contains 5611 articles and IT3 dataset contains 6651 articles with different classes. Remaining 2 datasets are the sports dataset names as Sports1 and sports2. The sports1 dataset contains 6777 number of articles while sports2 dataset contains 3936 number of articles. All these datasets are manually classified.

These all datasets are classified. The input to the system is given by the user which is called as Key term. This input can be a single or multiple terms. When the key term is given to the system, system compares the key term with the dataset to find out the related entity and we get the final output as a class which is semantically similar to the key term and its related articles.

Here, we have performed the serial and the parallel execution of the naive bayes algorithm. We are measuring the time for both type of executions. The time required for the serial execution of an algorithm is denoted by ts. While the time required for the parallel execution of an algorithm is denoted as tp. Finally we calculate the speed up in percent i.e Δt of the system with the following formula:

$$\Delta t = \frac{ts - tp}{ts} * 100 \qquad (2)$$

Table 6.1. The planning and control components.

Dataset	Key term	Ts	tp	Δt (%)	Related entity	Probabilistic scores
IT1	SQL	2min 32sec	21 sec	86.18	Mysql	0.2
IT2	Browser	4min 40sec	25 sec	91.42	Chrome	0.017
IT3	Apple	5min 32sec	23 sec	93.07	Macbook	0.04
IT1	Linux	2min 32sec	21 sec	86.18	Ubuntu	0.04
Sports1	Baseball	5min 39sec	24 sec	92.92	MLB	0.06
Sports2	Soccer	3min 20sec	31 sec	84.5	MLS	0.03

The table above shows that we are giving different key terms as an input to the different datasets. Accordingly, we are getting the time which is required for the serial and parallel execution of an algorithm. In addition, speed up is obtained which is calculated with the formula mentioned above. Due to parallelism speed up of 89% is obtained. We are getting the related entity of the key term as an output with its probabilistic score.

7. CONCLUSION

Here the proposed system, a naive bayes algorithm for semantic similarity measurements with parallelism is presented. The earlier method generates a vector of related Wikipedia entities as the semantic representation of a given text and uses the vector for measuring the semantic similarity. Whereas the proposed method is expected to aggregate the vectors using extended Naive Bayes (ENB). Our method generates refined results. Due to parallelism technique speed up of 89% is obtained.

REFERENCES

[1] S. Banerjee, K. Ramanathan, and A. Gupta, “Clustering Short Texts Using Wikipedia," in Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), July 2007, pp. 787-788.

[2] X. Sun, H. Wang, and Y. Yu, “Towards Effective Short Text Deep Classification," in Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), July 2011, pp. 1143-1144.

[3] Wikipedia, http://www.wikipedia.org/.

[4] P. Ferragina and U. Scaiella, “TAGME: On-the-y Annotation of Short Text Fragments (by Wikipedia Entities)," in Proceedings of ACM Conference on Information and Knowledge Management (CIKM), Oct. 2010, pp. 1625-1628.

[5] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, “Learning to classify short and sparse text Web with hidden topics from large-scale data collections," in Proc. 17th Int. World Wide Web Conf. (WWW), Apr. 2008, pp. 91-100.

[6] E. Gabrilovich and S. Markovitch, “Computing Semantic Relatedness using Wikipedia based Explicit Semantic Analysis," in Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), Jan. 2007, pp. 1606-1611.

[7] Masumi Shirakawa, Kotaro Nakayama, Takahiro Hara, Shojiro Nishio, “Wikipedia-based Semantic Similarity Measurements for Noisy Short Texts Using Extended Naive Bayes", IEEE Transactions On Emerging Topics In Computing, VOL. X, NO. 2168-6750 (c) 2015.

[8] E. Meij, W.Weerkamp, and M. de Rijke, “Adding semantics to microblog posts," in Proc. ACM Int. Conf. Web Search Data Mining (WSDM), Feb. 2012, pp. 563-572.

[9] X. Hu, N. Sun, C. Zhang, and T.-S. Chua, “Exploiting internal and external semantics for the clustering of short texts using world knowledge," in Proc. ACM Conf. Inf. Knowl. Manage. (CIKM), Nov. 2010, pp. 919- 928..

[10] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, “Short text conceptualization using a probabilistic knowledge base," in Proc. Int. Joint Conf. on Artif. Intell. (IJCAI), Jul. 2011, pp. 2330-2336.

[11] XiangWang, et al., “Short Text Classification using Wikipedia Concept based Document Representation", IEEE Transactions , 978-1-4799-2876-7/13

[12] PuWang et al., “Using Wikipedia knowledge to improve text classification", Published online: 17 September 2008 Springer-verlag London Limited 2008, Knowledge Information System (2009) 19:265-281, DOI 10.1007/s10115-008-0152-4.

[13] Pu Wang and Carlotta Domeniconi, “Building Semantic Kernels for Text Classification using Wikipedia" Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 713-721.

[14] M. Strube and S. P. Ponzetto, "WikiRelate! Computing Semantic Relatedness using Wikipedia," in Proceedings of National Conference on Artificial Intelligence (AAAI), July 2006, pp. 1419- 1424.

[15] C. Fellbaum, "WordNet: An Electronic Lexical Database," The MIT Press, May 1998

[16] S. P. Ponzetto and M. Strube, "Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution," in Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics(HLT-NAACL), June 2006, pp. 192-199.

[17] D. Milne and I. H. Witten, An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links," in Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI), July 2008, pp. 25-30.

[18] Y. Ollivier and P. Senellart, "Finding Related Pages Using Green Measures: An Illustration with Wikipedia," in Proceedings of National Conference on Artificial Intelligence (AAAI), July 2007, pp. 1427-1433.

[19] E. Yeh, D. Ramage, C. D. Manning, E. Agirre, and A. Soroa, "WikiWalk: Random Walks on Wikipedia for Semantic Relatedness," in Proceedings of Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4), Aug. 2009, pp. 41- 49.

[20] M. Ito, K. Nakayama, T. Hara, and S. Nishio, "Association Thesaurus Construction Methods based on Link Co-occurrence Analysis for Wikipedia," in Proceedings of ACM Conference on Information and Knowledge Management (CIKM), Oct. 2008, pp. 817-826.

[21] S. Hassan and R. Mihalcea, "Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge," in Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Aug. 2009, pp. 1192-1201.

[22] M. Yazdania and A. Popescu-Belis, "Computing Text Semantic Relatedness Using the Contents and Links of a Hypertext Encyclopedia," Artificial Intelligence, vol. 194, pp. 176- 202, Jan. 2013.

[23] M. A. H. Taieb, M. B. Aouicha, and A. B. Hamadou, "Computing Semantic Relatedness Using Wikipedia Features," Knowledge-Based Systems, vol. 50, pp. 260-278, Sept. 2013.

[24] K. Nakayama, T. Hara, and S. Nishio, "Wikipedia Mining for An Association Web Thesaurus Construction," in Proceedings of International Conference on Web Information Systems Engineering (WISE), Dec. 2007, pp. 322-334.

[25] Menaka S and Radha N, "Text Classification using Keyword Extraction Technique," Volume 3, Issue 12, December 2013.

[26] Abdullah Bawakid and Mourad Oussalah, "A Semantic-Based Text Classification System," IEEE 9th International Conference on Cybernetic Intelligent Systems, 01/2010; DOI: 10.1109/UKRICIS.2010.5898112.

BIOGRAPHIES

Sayali S. Rasane Is pursuing the Masters in Computer from G. E. S. R. H. Sapat College of Engineering., Nashik under Pune University. She has pursued his Bachelor's Degree in IT from N. D. M. V. P. College of Engineering., Nashik under Pune University.

Dipak V. Patil received B.E. degree in computer engineering in 1998 from University of North Maharashtra India and M. Tech. degree in computer Engineering in 2004 from Dr. B. A. Technological University, Lonere, India. He has done Ph.D. degree from S. R. T. M. University, Nanded. Currently he is an Associate Professor in Computer Engineering Department at G. E. S. R. H. Sapat College of Engineering., Nashik, India. His research interests are in data mining and soft computing.