The Observatorium – The structure of news: topic monitoring in online media with mutual information

David Rodrigues[1]

[1] ISCTE - Lisbon University Institute, gab. D609 – DCTI, Av. Forças Amadas, 1649-026 Lisboa, Portugal
david.rodrigues@gmail.com

Abstract: Large, real time text classification systems are becoming a popular topic. We present a method for automatically extracting correlated news from online media using a dynamic similarity graph and use the variation of information as a measure to identify topics, lifespan and key terms. The presented method has the advantage of requiring no human intervention or training and having no pre-assigned categories because they emerge from the dynamics of the generated network.

Keywords: text categorisation, term extraction, mutual information

1 Introduction

In recent years we have been witnessing a surge of interest in term extraction and automated text categorization[1-3], mainly because of the increasing amount of information that is being produced online. Examples include extracting and classifying biological text [4], the categorizing online news [5, 6] and personalized recommendation systems [7]. With online information's rapid growth has come the development of automated and agile methods to process it. The literature provides many examples of term extraction methods that can be categorized roughly into two fields, according to their different perspectives. One takes the task from a linguistic, terminology and natural language processing perspectives [8], and the other uses mainly tools from the statistical and information retrieval fields [2, 9].

One of the great challenges in the process of term extraction is related to the peculiarities of the language being processed. Traditional evaluation relied on assessments made by humans about extracted terms qualities. This evaluation method can be difficult to apply to large datasets. By combining several automated strategies we aim to reduce human intervention to the bare minimum.

Categorising text consists of assigning documents to a set of predefined categories (or labels). Several strategies have been proposed for this learning task, including, among others, regression models, Bayesian approaches, nearest neighbor classification, neural networks, hierarchical clustering and Support Vector Machines [1, 3, 10-13].

In this work we propose a method of automatically tackling the term extraction while simultaneously assigning the different documents to their own topics. The method uses variation of information, first proposed by Marina Meilã [14]. This measure is a metric and evaluates the distance between different clusterings. When coupled with the Jaccard index, the method provides a quick way of identifying sets of closely similar documents, from a dynamic network that is generated dynamically.

In the following section we discuss some of the related work done in this field. Subsequently we present the methodology employed at *The Observatorium* and explain the measures in use. In section 4 we present the some preliminary results and we conclude with an assessment of this approach goodness.

2 Related work: methods of categorization text

In recent years there interest has surged in the area of topic spotting, sometimes also called trend tracking. Researchers have applied an increasing number of learning approaches, including regression models, nearest neighbor classification, Bayesian probabilistic approaches, decision trees, inductive rule learning, neural networks, on-line learning and Support Vector Machines [1, 3, 10-13]. Most of those methods are supervised and require a training set where documents previously classified by humans are used as input to make the system learn each category's particular features. This approach poses two main problems: the need for a language-dependent analysis and classification by specialists and the difficulty finding new categories. A new text is either part of one of the existing categories or not a part of any of them at all. To solve these problems we use a dynamic network and variation of information [14].

Variation of information as a measure of topic change

We propose the use of *variation of information* (VI), which allows us to measure the amount of information lost and gained when changing from partitioning C to partitioning C' of the same data set [14]. The following paragraphs present the VI measure, synthesising a detailed description by Meilă [14]. Considering one partitioning C, the probability that a node k belongs to cluster C_k is given by equation (1) where n_k is the number of nodes in cluster C_k and n is the total number of its elements.

$$P(k) = \frac{n_k}{n} \tag{1}$$

The uncertainty associated with the measure is the entropy of the variable $P(k)$

$$H(C) = -\sum_{k=1}^{K} P(k)\log(P(k)) \tag{2}$$

where $H(C)$ is the entropy associated with the partitioning C. This is always a non-negative value, and is zero when there's no uncertainty. The mutual information

between two partitionings, C_k and C'_k , represented by $I(C_k, C'_k)$, means the information that one has over the other. The mutual information is given by the probability $P(k,k')$, representing the probability that a node belonging to the cluster C_k is in the cluster $C'_{k'}$.

$$P(k,k') = \frac{\left|C_k \cap C'_{k'}\right|}{n} \tag{3}$$

Using expression (3), the mutual information $I(C,C')$ is defined as the mutual information associated with the two random variables k and k':

$$I(C,C') = \sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k,k')\log\frac{P(k,k')}{P(k)P(k')} \tag{4}$$

Meilă [14] proposed VI as a way to compare clusters:

$$VI(C,C') = H(C) + H(C') - 2I(C,C') \tag{5}$$

This measure is a metric because it is always non-negative, it is symmetric and it presents triangular inequality.

3 Our approach at *The Observatorium*

In this paper we used the news gathered from the online Portuguese newspaper *Público* (http://publico.pt/) from November 7, 2009 to January 25 2010. We tracked the latest news RSS Feed and then downloaded the corresponding HTML files. We kept only the URL's HTML code, but stored no images or flash or binary objects.

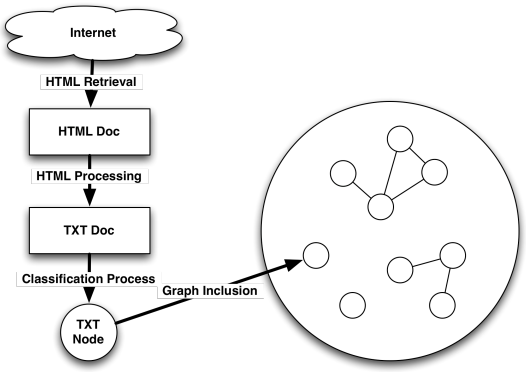

Figure 1 – Acquiring and processing text documents from the Internet

After the collection stage each HTML file has to be processed to remove duplicate files, HTML tags and artefact text. The problem with online media is that a webpage has textual information that isn't pertinent to the topic extraction phase. We also

observed that the pages structure isn't constant among all the retrieved pages, which prohibits the use of the HTML structure to retrieve the text as in Lin [15]. Newspaper journals usually include snippets of text from other stories and aren't therefore easy to identify. Also because the HTML pages include several navigational links, these anchors text shows up as isolated words. Publicity is also present and can pose problems.

To solve these issues, we extracted the text from each HTML file by employing the Text to Tag ratio proposed by Weninger [6] with a smooth factor of 1. This gives us a good elimination of the text artefacts discussed above. We did some filtering to eliminate words shorter than three characters and longer than 20 [16]. Some pre-processing techniques include other tasks, such as tokenization and stemming, but these are language-dependent. Here we aim to extract terms without previous knowledge of the texts language.

Our method adds each text document to a graph, G, that is dynamically generated from the received text nodes. We assumed each node to have a certain life expectancy in this graph, the time to live (TTL). We iteratively added these nodes to the graph and the Jaccard similarity (eq. 6) was computed between each node and all the previous nodes in the graph. Edges were then established from the added node to the previous nodes if their similarity was above a threshold j_{min}.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

Each time we added a node to graph G the TTL of the nodes that established connections to it is reset, but for others this value is decremented. This purges old text nodes that don't receive new connections after some time determined by the TTL.

At each time step we compute the VI from the previous graph to the new one $VI(G_{t-1}, G_t)$. From this value we can determine the points where VI exceeds a VI_{min} threshold. If the VI does exceed that threshold, it indicates the deletion of a reasonably large component of the graph. That also is the end of a topic in the timeline. From this information we can track the topic's origin by looking into that deleted component's oldest node, and the time span of each topic. We also can process these components via a text extraction algorithm based on term frequency or another similar technique (not described in this paper).

4 Results

The corpus of this analysis consisted of 7928 news items collected by Theseus[1] software from the *Público* newspaper from Nov 11, 2009 through Jan 25, 2010. We processed those items with TTL=100, j_{min}=0.5 and VI_{min} = 0.5. We obtained these parameters empirically from a coarse search and, at this moment, their influence on

[1] Theseus is a webpage retrieval system that works by parsing RSS feeds and collecting the corresponding webpages. It was developed internally as part of the *The Observatorium* project (http://theobservatorium.eu).

the methods quality is not fully asserted. These parameters are context dependent and can be adjusted online for each specific newspaper according to the results goodness.

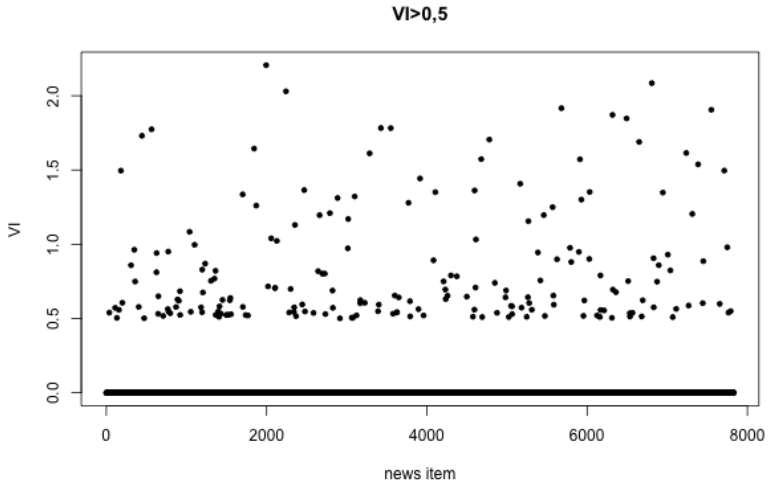

Figure 2 – VI evidence of topic deletion.

In Fig. 2 we present the VI between the graphs of two consecutive time steps that exceed 0,5. Because VI is a metric, the points that present higher values represent greater changes in the graph's structure than do points that present lower values, which usually meaning the removal of an entire cluster of news from the graph. As an example, at time step 2092 (VI=2,2), we had the removal of a cluster of 37 texts (from late November 2009) related to financial subjects. Those texts main subject was the health financial scandal in the USA in November 2009.

By applying the proposed method we identified 196 topics during the analysis period.

Table 1 – *Público* topic tracking results, av. lifespan and av. number news

n. topics	< Lifespan (/h)>	<Topic Size>
196	17.1	7.5

Those 196 topics had an average lifespan of 17.1 hours and the topic with the longest lifespan lived for 104.7 hours (approximately four days and nine hours). The results are in accordance with what we would expect from an online newspaper whose news concentrates mainly on daily topics with a few stories that "percolate" over several days.

5 Conclusions

The preliminary results show that a method for simultaneously and automatically extracting topics classifying text is possible. This method has the advantage of requiring no prior knowledge or training. The use of VI allows for a fast method, based on information theory, of processing large volumes of data and allows the combination of network theory in the process of discovering online media's news structure.

References

1. Cachopo, A.M.D.J.C., Oliveira, A.L.: An Empirical Comparison of Text Categorization Methods. String Processing and Information Retrieval (2003) 183-196
2. Pantel, P., Lin, D.: A Statistical Corpus-Based Term Extractor. Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence. Springer-Verlag (2001) 1-10
3. Yang, Y., A, X.L.: A re-examination of text categorization methods. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Vol. pages. ACM Press, New York, New York, USA (1999) 42-49
4. Lee, M., Wang, W., Yu, H.: Exploring supervised and unsupervised methods to detect topics in biomedical text. BMC bioinformatics 7 (2006) 140-140
5. Jo, T., Seo, J., Kim, H.: Topic Spotting on News Articles with Topic Repository by Controlled Indexing. Springer Berlin Heidelberg, Berlin, Heidelberg (2000)
6. Weninger, T., Hsu, W.H.: Text Extraction from the Web via Text-to-Tag Ratio. 2008 19th International Conference on Database and Expert Systems Applications (2008) 23-28
7. Zhang, Z.-k., Zhou, T., Zhang, Y.-c.: Personalized Recommendation via Integrated Diffusion on User-Item- Tag Tripartite Graphs. Physica A (2010) 179-186
8. Gravano, L.: An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering 1 Introduction 2 Clustering Algorithms 3 Linguistic Features. (1998) 224-231
9. Nigam, K., Lafferty, J., McCallum, A.: Using Maximum Entropy for Text Classification. IJCAI-99 Workshop on Machine Learning for Information Filtering (1999) 61-67
10. Miao, Y., Qiu, X.: Hierarchical Centroid-based Classifier for Large Scale Text Classification. (2010) 3-6
11. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Nédellec, C., Rouveirol, C. (eds.): European Conference on Machine Learning (ECML). Springer, Berlin (1998) 137-142
12. Hamamoto, M., Pan, J.-y.: A Comparative Study of Feature Vector-Based Topic Detection Schemes. Information Retrieval (2005)
13. Solé, R.V., Corominas-murtra, B., Valverde, S., Steels, L.U.C.: Language Networks: Their Structure, Function, and Evolution. Complexity 00 (2010) 1-7
14. Meilă, M.: Comparing clusterings-an information based distance. Journal of Multivariate Analysis 98 (2007) -895
15. Lin, S.-h., Ho, J.-m.: Discovering Informative Content Blocks from Web Documents. Knowledge Creation Diffusion Utilization (2002) 1-9
16. Cachopo, A.M.D.J.C.: Improving Methods for Single-label Text Categorization. Vol. PhD (2007) 167-167