Winnowing-Based Text Clustering

Javier Parapar
Information Retrieval Lab
Department of Computer Science
University of A Coruña
javierparapar@udc.es

Álvaro Barreiro
Information Retrieval Lab
Department of Computer Science
University of A Coruña
barreiro@udc.es

ABSTRACT

We present an approach to document clustering based on winnowing fingerprints that achieved good values of effectiveness with considerable save in memory space and computation time.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Clustering

General Terms: Algorithms, Experimentation, Performance.

Keywords: Document clustering, document representation, fingerprinting, experimentation.

1. INTRODUCTION

Document clustering has been demonstrated as a successful way of improving the performance of several tasks in Information Retrieval like document retrieval, text summarisation or results presentation. The main problem of applying clustering techniques in real retrieval systems is the computational cost.

Traditionally the clustering algorithms have a high computational complexity in terms of space and time [2]. Hierarchical methods are typically $O(n^2 \log(n))$ (n is the number of documents) and $O(n^2)$ in time and space respectively. K-means is $O(k \times i \times n)$ and $O(k + n)$ (k is the number of clusters, i the number of iterations), but has the problem that the quality of the clustering is quite reliant on the initial random cluster seeding among the collection. Meanwhile hybrid approaches as Buckshot are $O(k \times i \times n + n \log(n))$ and $O(ni)$, but these methods become inefficient in practice. So with large collections, e.g. the web, the needed time and memory space are not admissible. This has motivated a lot of research in two main ways: the design of cheaper clustering algorithms, and the use of techniques in order to reduce the input size: feature selection and dimensionality reduction.

Fingerprints of documents or *signatures* were used in information retrieval in several tasks including audio retrieval, image retrieval, and document retrieval. The typical uses of fingerprinting were detection of duplicates, identification of document plagiarism and in ad-hoc retrieval. Although data clustering using document fingerprints was not too much exploited, we want to remark two works. Broder et al. in [1] explored the use of a kind of fingerprints called *shingles* to perform syntactic clustering of web documents efficiently. Puppin and Silvestri [4] also evaluated the use of the *shingles* in order to get an efficient collection partition.

In this paper we propose the use of winnowing fingerprints [6] for the clustering task. Winnowing was presented with the objective of plagiarism detection, but the fingerprint construction guarantees also a set of theoretical properties in terms of fingerprint density and sub-string matching detection.

In order to compare the performance of the selected fingerprint method we chose three other document representations: term frequency, mutual information and a designed fixed-size fingerprint that we have coined as *n-fingerprint*: representations of the documents as n-gram frequency deviations from the standard frequency in a given language.

2. WINNOWING BASED CLUSTERING

One of the advantages of the winnowing algorithm presented in [6] for hashing selection is the trade-off between the fingerprint length and the shortest matching string to be detected, establishing theoretical guarantees.

Let t be the threshold that guarantees *(a)* that any match of strings *longer or equal than t is detected* , and let n be another threshold that guarantees *(b)* that any match of strings *shorter than n is not detected.* The parameters t and n are chosen by the user being the n value the n-gram size. Given any list of hashes $h_1, h_2, ...h_k$, if $k > t-n$ then at least one of the $h_{1<i<k}$ should be chosen in order to guarantee the detection of all matches equal or longer than t. To achieve this the next selection process was proposed:

Let $w = t - n + 1$ be the window size and let $h_1, h_2, ...h_k$ be the whole sequence of hashes result of hashing all the n-grams in which the document text was decomposed. Each position $1 \leq i \leq k - w + 1$ defines the start of a window (sub-list) of hashes $h_i, h_{i+1}, ...h_{i+w-1}$, therefore in order to guarantee the condition *(a)* is necessary and sufficient to select one hash value for every window to compose the fingerprint. The condition *(b)* is guaranteed by the fact of choosing n-grams of size n. In order to achieve this, the process defined in [6] was:

In each window select the minimum hash value. If there is more than one hash with the minimum value select the rightmost occurrence. Now save all selected hashes as the fingerprints of the document.

After applying the winnowing algorithm each document is a multi-set of hash values. Each multi-set $W(d)$ has different size depending on the document d size and could have repeated hash values. In order to cluster the collection we have to adopt a suitable similarity measure. For this pur-

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

pose we have adapted the Jaccard Coefficient computed as in equation 1 to multi-sets.

$$Sim(d_i, dj) = \frac{|W(d_i) \cap W(d_j)|}{|W(d_i) \cup W(d_j)|} \quad (1)$$

The union was defined as: $W(d_i) \cup W(d_j)$ is the multi-set composed of every element of $W(d_i)$ and $W(d_j)$ repeated as many times as its maximum presence in $W(d_i)$ or $W(d_j)$. The intesection was defined as: $W(d_i) \cap W(d_j)$ is the multi-set composed of those hashes present in $W(d_i)$ and $W(d_j)$ repeated as many times as in the multi-set where they have lower cardinality. For further reference we shall also define here the join of $W(d_i) \uplus W(d_j)$ as the multi-set composed of all the elements of $W(d_i)$ and all the elements of $W(d_j)$.

Another important point in the case of the clustering methods that deal with centroids is the centroid computation. When the documents are represented with term frequency, mutual information or n-fingerprints the cluster centroid is computed as the average value for every term in the vector for all the documents of the cluster. In our case this approach does not make sense.

The problem of computing cluster centroids for sets has previously been addressed [7] and it is not trivial. One alternative could be using the union of the cluster documents as centroid, but it will have a very high cardinality and will not consider the frequency. Thus, when the cluster contains documents with very few elements in common the centroid will not be representative. On the other hand, intersection is not suitable neither because it is very probable to obtain an empty set when working with several documents. The same problems would happen when working with multi-sets, so we have devised a new approach to centroid computation with multi-sets in order to avoid them.

For each cluster $C = \{W(d_1), ...W(d_n)\}$ the multi-set representing its centroid $centroid(C)$ is computed as in eq. 2.

$$centroid_\gamma(C) = \bigcup f(C_\gamma, h) \mid h \in \biguplus_{i=1..n} W(d_i), \frac{hf(h,C)}{n} \geq \gamma \quad (2)$$

Where $hf(h, C)$ is the number of occurrences of hash h in the join of all the documents of the cluster C, and $f(C_\gamma, h)$ is a function that for a hash h returns a multi-set composed of h repeated so many times as $(\frac{hf(h,C)}{n})/\gamma$.

3. EVALUATION AND RESULTS

In order to asses the outcomes of clustering based on winnowing fingerprints, evaluation of the efficiency and effectiveness was carried out. The comparison was made between the use of the whole documents with different representations: term frequency (TF) and mutual information (MI); n-fingerprints (NFP) with n=1 and winnowing fingerprints (WFP) with $w = 25$, n=5 and $\gamma = 0.3$.

3.1 Test Data

From Reuters-21578 V.12 collection, which is divided in 92 not disjointed classes, the 2745 documents that were assigned at least to one topic, had the attribute *"LEWISSPLIT=TEST"* and had the *BODY* element were selected.

3.2 Effectiveness Measures

The effectiveness of the clustering algorithms was assessed using four different metrics. All the metrics are based on comparing the clustering outcomes with another manually done split (answer key that defines the classes) of the collection that is used as a judgement criteria. Tree edit was defined as in [3] and purity, F-measure and entropy were defined as in [5].

3.3 Results

Algorithm		TF	MI	WFP	NFP
K-means	TE:	0.60	0.61	0.57	0.46
	P:	0.78	0.72	0.69	0.59
	F:	0.33	0.29	0.34	0.16
	E:	0.011	0.015	0.014	0.013
	Times:	42.4s	25.6s	18.3s	10.4s
Average-link	TE:	0.53	0.58	0.56	0.38
	P:	0.77	0.80	0.69	0.57
	F:	0.73	0.69	0.63	0.40
	E:	0.010	0.012	0.013	0.018
	Times:	2441.7s	643.8s	367.1s	293.6s

Table 1: Results over the Reuters Collection (K=92)

Attending to these results (see Table 1) the first fact to remark is that the results of using winnowing fingerprint are better than the results using n-fingerprint. With average-link for the tree edit distance the winnowing fingerprint even outperformed the term frequency representation and also with the k-means algorithm and with the F-measure. Comparing winnowing fingerprint with mutual information representation we obtained that MI clearly outperforms WFP and also TF; this is explained by the fact that it uses more information and collection statistics for its computation. This also produces that the computation of MI vectors is slower than TF and even more than WFP.

4. CONCLUSIONS AND FUTURE WORK

Based on the obtained results the use of this kind of techniques resulted acceptable in terms of trade-off between effectiveness and efficiency. We have to perform further evaluation over larger collections. We also want to test how the method can be applied to other document representations like summaries or snippets. Another objective is to evaluate indirectly this clustering approach is some of the mentioned task.

Acknowledgements: This work was co-funded by FEDER, SEUI and Xunta de Galicia under projects TIN2005-08525-C02 and 07SIN005206PR

5. REFERENCES

[1] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. In *WWW '97*, pages 1157–1166, 1997.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[3] P. Pantel and D. Lin. Document clustering with committees. In *SIGIR '02*, pages 199–206,2002.

[4] D. Puppin and F. Silvestri. The query-vector document model. In *CIKM '06*, pages 880–881, 2006.

[5] M. Rosell, V. Kann, and J.-E. Litton. Comparing comparisons: Document clustering evaluation using two manual classifications. In *ICON'04*, 2004.

[6] S. Schleimer, D. S. Wilkerson, and A. Aiken. Winnowing: local algorithms for document fingerprinting. In *SIGMOD '03*, pages 76–85, 2003.

[7] F. Giannotti and C. Gozzi. Characterizing web user accesses: A transactional approach to web log clustering in *ITCC '02*, pages 312–317, 2002.