Defining the Dynamicity and Diversity of Text Collections

Ilya Markov and Fabio Crestani

University of Lugano, Faculty of Informatics
Via G. Buffi 13, 6900, Lugano, Switzerland
{ilya.markov,fabio.crestani}@usi.ch

Abstract. In Information Retrieval collections are often considered to be relatively dynamic or diverse, but no general definition has been given for these notions and no actual measure has been proposed to quantify them. We give intuitive definitions of the dynamicity and diversity properties of text collections and present measures for calculating them based on the notion of novelty. Experimental results show that the proposed measures are consistent with the definitions and can distinguish collections effectively according to their dynamicity and diversity properties.

1 Introduction

Each time some Information Retrieval technique is extensively tested, researchers try to consider several experimental collections that are different one from the other according to a number of properties. For example, Federated Search testbeds are explicitly said to be homogeneous or heterogeneous [2], special smoothing techniques are to be applied for homogeneous collections [4], query expansion may be performed based on the terms from recent documents in relatively dynamic collections. Therefore it is important to define what the dynamicity[1] and diversity of text collections are and to be able to measure these properties quantitatively.

We consider text collections that evolve over time. This assumption comes from the real-world environment, where the documents are added to a collection or updated as time goes by (we do not consider document removal since it happens rarely nowadays). Our intuition behind measuring the *dynamicity property* of text collections is as follows. Given an evolving collection, if a newly added document is novel comparing to the documents added in the nearest past, then the collection can be considered to be relatively dynamic as opposed to a collection for which new documents are redundant comparing to the documents seen in the nearest past.

In order to measure the *diversity property* one has to compare all the documents in a collection pairwise. Based on this, one collection can be said to be more diverse than the other if, in general, its documents are more different between each other, then the documents of the other collection. Only some subset of documents in a collection can be chosen for pairwise comparison to reduce the computational cost.

[1] In the context of this paper dynamicity means the property of being dynamic.

M. Lalmas et al. (Eds.): ECDL 2010, LNCS 6273, pp. 474–477, 2010.

2 Related Work

In the past years a number of research areas have been concerned with improving their results by incorporating diversity: text document retrieval [3], recommender systems [7], image retrieval [5]. As opposed to these works, our intent is to find standalone measures for estimating the diversity and dynamicity properties of text collections in general. As far as we know nobody tried to quantitatively estimate the dynamicity property of text collections.

Our intuition behind measuring the dynamicity and diversity properties of text collections is based on the notion of *document novelty*. First, the task of identifying novel and redundant documents was addressed by Zhang et al. [6]. A number of measures were proposed based on three different types of evidence: new word counts, cosine distance and distribution similarity (Kullback-Leibler divergence). It was shown that cosine distance performs the best in the task of novelty and redundancy detection for AP and WSJ TREC datasets.

Allan et al. [1] applied the measures discussed in [6] to the task of novelty detection at a sentence level. It was shown that on a sentence level word counting measures perform the best. This is because sentences with overlapping vocabulary but different word distributions are considered novel from a word distribution point of view, but redundant from a new word counting perspective.

In this work we do not consider the task of document novelty and redundancy detection itself, but apply the measures discussed in [6] and [1] to the task of calculating the dynamicity and diversity measures of text collections.

3 Measures

Each *novelty measure* receives a document d and a set of documents DS ($d \notin DS$) as an input and returns the novelty score of a given document against a given set of documents $NS(d, DS)$. The novelty measures we use include:

Average New Word Ratio: $AvgNWR(d, DS) = \frac{\sum_{d_i \in DS} |W_d \cap \overline{W_{d_i}}|}{|W_d| \cdot |DS|}$. Here W_d is a set of words in a document d.

Cosine Distance: $CosDist(d, DS) = \frac{\sum_{d_i \in DS} cos(d, d_i)}{|DS|}$. Here V is a combined vocabulary of a document d and a set of documents DS. A document d is represented as a vector $d = (w_{1,d}, w_{2,d}, ..., w_{n,d})^T$ and $cos(d_1, d_2) = \frac{\sum_{t \in V} w_{t,d_1} w_{t,d_2}}{\|d_1\| \cdot \|d_2\|}$.

Average Language Model KL Divergence: $AvgLM(d, DS) = \frac{\sum_{d_i \in DS} KL(\Theta_d, \Theta_{d_i})}{|DS|}$. Here $KL(\Theta_d, \Theta_{d_i}) = \sum_{t \in V} p(w|\Theta_d) log \frac{p(w|\Theta_d)}{p(w|\Theta_{d_i})}$ and Jelinek-Mercer smoothing is used.

In order to calculate the *dynamicity property* of a text collection new documents of a collection are compared against the previous documents of the same collection. Documents added in the nearest past form a set of documents DS of size N, thus DS can be viewed as a window of size N. As a new document d arrives, its novelty score against DS is calculated. Then d is added to DS,

while the oldest document in DS is removed (i.e. the window moves one step forward preserving only the last N documents). Thus the final dynamicity score is computed as follows.

Input: Collection C of documents sorted chronologically; set of documents DS formed out of first N documents in C
Output: Sum of novelty scores of all documents in C
foreach *document* $d \in C \setminus DS$ **do**
  $DynScore(C) += NS(d, DS)$;
  $DS = (DS \cup \{d\}) \setminus \{d_{oldest}\}$;
end

The final dynamicity score is normalized by the number of documents in a collection: $DynScore(C) = \frac{DynScore(C)}{|C|}$. Window size N is a parameter that can be chosen according to a collection statistics.

In order to measure the *diversity property* of a text collection we choose a random sample of documents RS of size M and compare these documents pairwise, since there is no notion of time in this case. Therefore the diversity score is calculated as follows: $DivScore(C) = \frac{\sum_{d \in RS} NS(d, RS \setminus \{d\})}{|RS|}$. Here M is a parameter that can be chosen according to a collection statistics.

4 Experiments

To test the proposed dynamicity and diversity measures we choose several intuitively different collections from TREC volumes 1-3, namely AP, WSJ, FR and Patents datasets. We expect the collections of news articles (AP and WSJ) to be ranked higher than FR and Patents according to both dynamicity and diversity properties, although it is not obvious which one of FR and Patents collections is more dynamic or more diverse.

The results for the dynamicity and diversity measures are presented in table 1. Due to the space limitations only the dynamicity measures will be discussed. The values of the diversity measures are calculated in a similar way, therefore the following discussion is applicable to them as well.

The ranking of the collections according to their dynamicity property is the same for $N = 10$ and $N = 100$, therefore the proposed dynamicity measures

Table 1. Dynamicity and diversity measures

Coll.	Dynamicity						Diversity					
	AvgNWR		CosDistTF		AvgLM		AvgNWR		CosDistTF		AvgLM	
	$N = 10$	$N = 100$	$N = 10$	$N = 100$	$N = 10$	$N = 100$	$M = 100$	$M = 1000$	$M = 100$	$M = 1000$	$M = 100$	$M = 1000$
WSJ	0.84	0.85	0.85	0.86	2.18	2.71	0.85	0.85	0.85	0.86	2.74	2.91
AP	0.84	0.85	0.81	0.83	1.98	2.56	0.86	0.86	0.83	0.83	2.63	2.81
Patents	0.65	0.65	0.73	0.74	1.62	1.94	0.66	0.66	0.74	0.74	1.97	2.06
FR	0.67	0.72	0.41	0.45	1.41	1.89	0.73	0.72	0.45	0.45	1.92	1.98

are stable across different window sizes N. As we expected, all the measures clearly split the collections in question into two sets: more dynamic collections of news articles (AP and WSJ) and less dynamic FR and Patents collections. However, there is no clear ranking inside those sets: AvgNWR-based measure give higher dynamicity score to the AP collection compared to the WSJ and to the FR collection compared to the Patents. CosDistTF- and AvgLM-based measures, on the other hand, produce the reverse ranking: the WSJ collection is considered to be more dynamic than the AP and the Patents collection to be more dynamic than the FR. We are currently investigating these results.

5 Future Work

There are a number of possible applications of the proposed measures. In Distributed Information Retrieval resource description, resource selection and results fusion algorithms may benefit from the knowledge of the values of dynamicity and diversity properties of federated collections. Special smoothing techniques could be applied for collections known to be relatively homogeneous (i.e. not diverse). Also query expansion may use only terms from recent documents in relatively dynamic collections. We are going to address all these questions in future work.

References

1. Allan, J., Wade, C., Bolivar, A.: Retrieval and novelty detection at the sentence level. In: Proc. of the ACM SIGIR. pp. 314–321. ACM (2003)
2. Callan, J.: Advances in Information Retrieval, chap. 5. Distributed Information Retrieval, pp. 127–150. Kluwer Academic Publishers (2000)
3. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proc. of the ACM SIGIR. pp. 335–336. ACM (1998)
4. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proc. of the ACM SIGIR. pp. 275–281. ACM (1998)
5. Song, K., Tian, Y., Gao, W., Huang, T.: Diversifying the image retrieval results. In: Proc. of the ACM MM. pp. 707–710. ACM (2006)
6. Zhang, Y., Callan, J., Minka, T.: Novelty and redundancy detection in adaptive filtering. In: Proc. of the ACM SIGIR. pp. 81–88. ACM (2002)
7. Ziegler, C.N., McNee, S.M., Konstan, J.A., Lausen, G.: Improving recommendation lists through topic diversification. In: Proc. of the WWW. pp. 22–32. ACM (2005)