DeShaTo: Describing the Shape of Cumulative Topic Distributions to Rank Retrieval Systems without Relevance Judgments

Radu Tudor Ionescu[1], Adrian-Gabriel Chifu[2], and Josiane Mothe[3]

[1] Faculty of Mathematics and Computer Science, University of Bucharest, Romania
raducu.ionescu@gmail.com
[2] IRIT UMR5505, CNRS, Université de Toulouse, Université Paul Sabatier, France
adrian.chifu@irit.fr
[3] IRIT UMR5505, CNRS, Université de Toulouse, ESPE, France
josiane.mothe@irit.fr

Abstract. This paper investigates an approach for estimating the effectiveness of any IR system. The approach is based on the idea that a set of documents retrieved for a specific query is highly relevant if there are only a small number of predominant topics in the retrieved documents. The proposed approach is to determine the topic probability distribution of each document offline, using Latent Dirichlet Allocation. Then, for a retrieved set of documents, a set of probability distribution shape descriptors, namely the skewness and the kurtosis, are used to compute a score based on the shape of the cumulative topic distribution of the respective set of documents. The proposed model is termed *DeShaTo*, which is short for *De*scribing the *Sha*pe of cumulative *To*pic distributions. In this work, DeShaTo is used to rank retrieval systems without relevance judgments. In most cases, the empirical results are better than the state of the art approach. Compared to other approaches, DeShaTo works independently for each system. Therefore, it remains reliable even when there are less systems to be ranked by relevance.

Key words: information retrieval, topic modeling, LDA, document topic distribution, skewness, kurtosis, ranking retrieval systems.

1 Introduction

Automatically estimating the effectiveness of any information retrieval system is one of the most important tasks in information retrieval (IR). An approach that could solve this task with a high degree of accuracy would have a broad range of applications including selective IR, selective query expansion [4, 15], ranking retrieval systems without relevance judgments [9,12,13], query difficulty prediction [3,11], to name only a few. Being able to understand and distinguish the behavior of a highly effective IR system from a poorly effective one (on a per query basis) is the key in solving the task of estimating IR effectiveness. Intuitively, a highly effective system should return a set of documents in which

there are only one or a few predominant topics[4] (related to the query), while an average or poorly performing system will return documents from various topics, since not all the documents will be relevant for the given query. In other words, more topics indicate that the given query is more ambiguous from the point of view of an IR system. Interestingly, this hypothesis represents the cornerstone of the clarity score [3], but there are many other aspects of relevance that are ignored by this supposition. Nevertheless, the same hypothesis is explored into a different direction in this work. More precisely, the current work proposes an approach that can potentially be used for estimating the relevance level of any IR system. Latent Dirichlet Allocation (LDA) [2] is employed to model the topics within a document collection offline. Then, by describing the shape of the cumulative topic distribution generated by the top documents retrieved by an IR system for a given query, it can be easily determined if the behavior of the respective system resembles the behavior of a highly or rather poorly effective system. Therefore, the proposed approach is termed *DeShaTo*, which is short for *De*scribing the *Sha*pe of cumulative *To*pic distributions. Finally, the proposed approach computes a score based on a combination of two probability distribution shape descriptors, namely skewness and kurtosis [5], which are computed on the cumulative topic distribution of the retrieved documents.

A series of experiments are conducted to validate the underlying hypothesis of the DeShaTo approach. Relevant document sets are tested against non-relevant document sets for all the queries available in the TREC Robust Track, Web Track 2013 and Web Track 2014 collections. In almost 95% of the cases, the DeShaTo approach is able to identify which set of documents is relevant, proving that the underlying hypothesis holds in most cases. Next, the DeShaTo score is averaged on the queries of each data set to produce rankings of the retrieval systems submitted for the respective TREC tracks. The DeShaTo approach is compared with a state of the art approach for the task of ranking retrieval systems without relevance judgments, namely *nruns* [9], using the Kendall Tau correlation measure. The results presented in [9] indicate that nruns is more accurate compared to the previous works [1, 12, 13]. Therefore, DeShaTo is only compared with nruns in the experiments. The overall empirical results presented in this work indicate that the DeShaTo approach is able to obtain a higher correlation with the true Average Precision (AP) scores.

Unlike most approaches for ranking retrieval systems [10, 13], including the state of the art approach [9], DeShaTo does not require information about other retrieval systems when dealing with one system. Indeed, nruns [9] is based on sharing information among systems to produce a set of pseudo-relevant documents for each query, while DeShaTo works independently for each system and thus, it can produce accurate results when there are less systems to be ranked.

[4] *Topic* represents here the theme of a text, as in topic modeling. In IR evaluation programs such as TREC, a *topic* refers to the information need. To avoid any confusion with the LDA topics, TREC topics are referred to as *queries* throughout this paper, therefore *query* can mean either the information need or to the text submitted to the search engine.

Some approaches, such as [8], require human assessments, while DeShaTo does not involve human effort. Another distinctive trait of DeShaTo is the employment of topic modeling for ranking retrieval systems. Moreover, DeShaTo is a general approach with high potential for other applications such as query difficulty prediction, selective query expansion and selective IR.

The rest of the paper is organized as follows. The DeShaTo approach is presented in Section 2. The validation and the experiments are described in Section 3. The final remarks are drawn in Section 4.

2 Describing the Shape of Cumulative Topic Distributions

The DeShaTo approach is based on the hypothesis that if there are more predominant topics that emerge in the set of documents retrieved for a given query, then the system effectiveness for the respective query is lower. On the other hand, if there are less predominant topics in the set of documents, then it means that the system is highly effective. Naturally, the topic distributions of the retrieved documents have to be computed in order to determine the effectiveness level of the IR system. In this work, Latent Dirichlet Allocation based on Gibbs sampling [2] is employed to compute the topic distributions of the documents, but other topic modeling approaches could possibly work equally well or even better [6]. Nevertheless, it is worth mentioning that LDA has successfully been used in different contexts in IR [7, 14].

Although DeShaTo is a post-retrieval approach, the topic distribution can be computed offline, right after indexing the documents, in order to reduce the online processing time. If LDA is carried out offline, the topic distributions can be immediately retrieved along with the documents when necessary.

In order to obtain a unique representation from all the topic distributions associated to the top retrieved documents for a query, the distributions have to be somehow combined into a single distribution. Instead of choosing only one way of cumulateing the topic distributions, three alternative ways are simultaneously employed, namely, the component-wise sum, the component-wise minimum, and the component-wise product, respectively. It is important to note that the three cumulative distributions have to be normalized, such that they all remain probability distributions (the sum of all the components is 1). The sum, the minimum and the product produce slightly different cumulative distributions and using them all together provides useful information for the next step. The relevant documents set contains only one or two predominant topics, while the non-relevant documents produce a mixture of predominant topics. What remains to be done from this point on is to find a way of comprising this difference in a measure or score. More formally, the next step is to find a robust approach to describe the shape of the cumulative topic distributions. The proposed approach uses two probability distribution shape descriptors, namely skewness and kurtosis [5]. More common statistics such as the mean or the standard deviation have also been tested out, but they have been found to be less informative. In probability theory and statistics, *skewness* is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The

skewness can be computed as the third central moment of the input probability distribution, divided by the cube of its standard deviation. In a similar way to the skewness, *kurtosis* is a descriptor of the shape of a probability distribution. More precisely, the kurtosis quantifies the peakness (width of peak) of the probability distribution of a real-valued random variable. The kurtosis can be computed as the fourth central moment of the probability distribution, divided by fourth power of its standard deviation.

High values of skewness and kurtosis indicate that the topic distribution is characterized by a small number of predominant topics, while low values of these statistics indicate that there are more (or even no) predominant topics. Therefore, the two statistics reflect exactly what is required to determine if the system behavior is good or poor with respect to a given query. Finally, the skewness and the kurtosis are embedded in the DeShaTo score that can be computed using the following closed form equation:

$$score = k(S) + s(S) + k(M) + s(M) + k(P) + s(P), \tag{1}$$

where k and s are two functions that return the kurtosis and the skewness of a probability distribution given as parameter, and S, M and P are the cumulative topic distributions obtained by computing the sum, the minimum and the product of the topic distributions corresponding to the retrieved documents. The probability distribution shape descriptors are combined in a very natural straightforward manner in Equation (1). By trying various combination schemes, a more efficient way of combining these descriptors can supposedly be found. For instance, a weighted sum could probably work better in practice, if the weights are learned on some training data. However, adding more parameters to DeShaTo is not necessarily desirable. Proposing alternative combination schemes will be properly addressed in future work.

3 Experiments

3.1 Data Sets Description

The TREC collections[5] that are being used in the experiments are presented next. They contain a set of information need statements, the document set and the relevance judgments for each query. The experiments are conducted on precisely three data sets, namely Robust, TREC Web Track 2013 and TREC Web Track 2014.

The results provided by participants are termed runs. TREC evaluates the runs using various effectiveness measures. The participant runs can thus be ranked according to one of these measures, such as the Mean Average Precision (MAP) over queries. In the experiments, all the queries along with all the participant runs from Robust, Web Track 2013 and Web Track 2014 are used to rank retrieval systems without relevance judgments. A summary of the data used in the experiments is given in Table 1.

[5] http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=10667

Table 1. A summary of the data sets used for the task of ranking retrieval systems without relevance judgments.

Data Set	Query IDs	# Queries	# Systems
Robust	$301-450, 601-700$	249	110
Web Track 2013	$201-250$	50	61
Web Track 2014	$251-300$	50	30

Table 2. Accuracy of the DeShaTo score for correctly identifying the set of relevant documents tested against a set of non-relevant documents. The number of LDA topics is 100.

Data Set	# Queries	# Correct Predictions	Accuracy
Robust	249	231	92.77%
Web Track 2013	50	50	100%
Web Track 2014	50	50	100%
Overall	349	331	94.84%

The DeShaTo approach is compared with a state of the art approach, namely nruns [9], using the Kendall Tau correlation measure, as in [9]. To reduce the offline processing time of DeShaTo, only the documents retrieved in the participant runs where included in the topic modeling process.

3.2 Empirical Validation of the Hypothesis

To validate the underlying hypothesis of DeShaTo, a simple procedure has been designed as described next. For each query in the three collections, a set of 30 relevant documents is produced by randomly sampling the documents. Likewise, another set of 30 non-relevant documents is produced for each query. For each query, 20 draws were made to randomly select the documents within the relevant and the non-relevant sets, in order to reduce the result of chance. Remarkably, the results of different trials are consistent with each other.

The cornerstone hypothesis of this work can be validated if it can identify, with a high degree of accuracy, which is the set of relevant documents only by using the DeShaTo score proposed in Equation (1). As such, the DeShaTo score was put to the test and the results are presented in Table 2. The DeShaTo score seems to be able to make good predictions in most of the cases. Indeed, there are only 18 queries from the Robust collection for which the score associated to the non-relevant document set is higher than the score associated to the relevant document set. The accuracy goes up to 100% for the Web Track 2013 and 2014 data sets. Overall, the accuracy of the DeShaTo approach is 94.84%. Although not perfect, this result offers some empirical proof that the underlying hypothesis of DeShaTo works well enough in practice.

3.3 Parameter Tuning

The DeShaTo approach is based on the topics modeled by LDA, but the number of topics could influence the accuracy of the proposed approach. Wei and Croft also observed that the number of topics affects the retrieval performance [14].

Table 3. Kendall Tau correlation between the ground truth ranking according to the MAP@30 measure and the systems ranking determined by the DeShaTo score. The best correlation per data set is highlighted in bold.

Data Set	50 **Topics**	100 **Topics**	250 **Topics**
Robust	**0.4286**	0.3524	0.3143
Web Track 2014	0.1190	0.1905	**0.3048**

Therefore, the number of topics is tuned on the Robust and the Web Track 2014 data sets through a validation procedure. Since the documents from the Web Track 2013 and 2014 data sets are from ClueWeb12, the number of topics validated on Web Track 2014 is also used on the Web Track 2013 collection. From each data set, 10% of the queries and 15 systems are chosen at regular interval. More precisely, one in every 10 queries are used for validation. One in every 7 systems are used for the Robust data set, while for the Web Track 2014, one in every two systems are used for validation. The amount of observations ($\#queries \times \#systems$) used for validation is deliberately chosen such that it is significantly smaller than the total amount of data, in order to prevent any kind of overfitting. Only 1.37% of the Robust data is used for validation. In a similar manner, 1.65% of the Web Track 2013 and 2014 data is used for tuning the number of topics.

The validation procedure aims to choose between using 50, 100 or 250 topics by evaluating the Kendall Tau correlation between the systems ranking determined by the DeShaTo score and the ground truth ranking determined by the Average Precision of the top 30 retrieved documents per run, namely AP@30. Actually, to produce the rankings, the score of each system has to be averaged over all the validation queries. Thus, the ground truth rankings are given by the Mean Average Precision of the top 30 documents, namely MAP@30. An interesting remark is that very similar results are obtained using the top 10 or the top 100 retrieved documents, but since nruns [9] was evaluated using the top 30 documents, the results presented in Table 3 and throughout this paper are also based on the top 30 documents per run. According to the best Kendall Tau correlations reported in Table 3, 50 topics will be used when LDA is carried out on the Robust documents. On the other hand, 250 topics will be used when LDA is carried out on the Web Track 2013 and 2014 documents. This difference can probably be explained by the type of documents that constitute the collections. The documents within the Robust collection are quite homogeneous since they are extracted from newspapers, while the documents within ClueWeb12 are web documents. In the latter collection, documents are much more heterogeneous and may contain topics that are not related to the document content such as links to the home page, menu buttons and so on.

3.4 Ranking Retrieval Systems Results

The DeShaTo score is compared with nruns [9] for the task of ranking retrieval systems without relevant judgements and the results are presented in Table 4. The values given in Table 4 represent the Kendall Tau correlations between the

Table 4. Kendall Tau correlation between the ground truth ranking according to the MAP@30 measure and the systems ranking determined by the DeShaTo score, on one hand, and the systems ranking produced by nruns, on the other. The best correlation per data set is highlighted in bold.

Data Set	**nruns [9]**	**DeShaTo**
Robust	**0.6195**	0.6112
Web Track 2013	0.1005	**0.2306**
Web Track 2014	0.4529	**0.4966**

ground truth ranking given by the MAP@30 values and the systems ranking determined by the DeShaTo score, on one hand, and by nruns, on the other hand. It is important to mention that the correlation reported for nruns in this paper (0.6195) is lower than the correlation reported in [9] (0.640). The difference comes from the fact that here the correlation is based on 249 queries from 2004 and 2005, instead of only the 150 queries from 2004. Furthermore, the correlation is here computed with respect to the MAP@30 score instead of the MAP score as in [9], which is actually more fair, since the predictions are made on the top 30 documents per query. Compared to nruns, the DeShaTo score gives a higher correlation for the Web Track 2013 and 2014 participant runs, while it produces only a slightly lower correlation for the Robust runs. This could be explained by the fact that nruns becomes unreliable for a small set of runs, because it leverages the information from multiple systems to produce a good set of pseudo-relevant documents. Web Track 2013 and 2014 have considerably less participants than Robust, and the nruns approach is less accurate on the Web Track 2013 and 2014 data sets. Unlike nruns, the DeShaTo approach relies solely on the results of a system to compute its score, which seems to be an advantage for the newer TREC collections. The overall results seem to indicate that DeShaTo is better than nruns.

4 Conclusion

This paper presented an approach that is able to distinguish between a highly effective IR system and a less effective IR system for some queries. The proposed approach is based on *De*scribing the *Sha*pe of cumulative *To*pic distributions modeled by LDA, hence the name DeShaTo. A set of experiments have been conducted in order to validate the underlying hypothesis of DeShaTo in practice. Moreover, another set of experiments have been conducted to compare DeShaTo with nruns [9] for the task of ranking retrieval systems without relevance judgments. The results indicate that DeShaTo gives a higher correlation with the MAP@30 score in most cases, most likely because its accuracy does not depend on the number of systems used. The described approach does not take into account the query text itself, but this will be covered in future work by analyzing the topic distribution of the query in relation to the document distributions.

References

1. Aslam, J.A., Savell, R.: On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. pp. 361–362 (2003)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting Query Performance. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 299–306 (2002)
4. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: A Framework for Selective Query Expansion. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. pp. 236–237 (2004)
5. Groeneveld, R.A., Meeden, G.: Measuring Skewness and Kurtosis. Journal of the Royal Statistical Society. Series D (The Statistician) 33(4), 391–399 (1984)
6. Mimno, D.M., McCallum, A.: Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. In: Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence. pp. 411–418 (2008)
7. Park, L., Ramamohanarao, K.: The Sensitivity of Latent Dirichlet Allocation for Information Retrieval. In: Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, vol. 5782, pp. 176–188 (2009)
8. Pavlu, V., Rajput, S., Golbus, P.B., Aslam, J.A.: IR System Evaluation Using Nugget-based Test Collections. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. pp. 393–402 (2012)
9. Sakai, T., Lin, C.Y.: Ranking Retrieval Systems without Relevance Assessments — Revisited. In: The 3rd International Workshop on Evaluating Information Access (EVIA) - A Satellite Workshop of NTCIR-8. pp. 25–33 (2010)
10. Shi, Z., Wang, B., Li, P., Shi, Z.: Using Global Statistics to Rank Retrieval Systems without Relevance Judgments. In: Intelligent Information Processing V, vol. 340, pp. 183–192 (2010)
11. Shtok, A., Kurland, O., Carmel, D., Raiber, F., Markovits, G.: Predicting Query Performance by Query-Drift Estimation. ACM Transactions on Information Systems 30(2), 11:1–11:35 (May 2012)
12. Soboroff, I., Nicholas, C., Cahan, P.: Ranking Retrieval Systems Without Relevance Judgments. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 66–73 (2001)
13. Spoerri, A.: Using the structure of overlap between search results to rank retrieval systems without relevance judgments. Information Processing & Management 43(4), 1059–1070 (2007)
14. Wei, X., Croft, W.B.: LDA-based Document Models for Ad-hoc Retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 178–185 (2006)
15. Winaver, M., Kurland, O., Domshlak, C.: Towards Robust Query Expansion: Model Selection in the Language Modeling Framework. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 729–730 (2007)