Modeling Score Distributions

M

Anca Doloc-Mihu
University of Louisiana at Lafayette, USA

INTRODUCTION

The goal of a web-based retrieval system is to find data items that meet a user's request as fast and accurately as possible. Such a search engine finds items relevant to the user's query by scoring and ranking each item in the database. Swets (1963) proposed to model the distributions of these scores to find an optimal threshold for separating relevant from non-relevant items. Since then, researchers suggested several different score distribution models, which offer elegant solutions to improve the effectiveness and efficiency of different components of search systems.

Recent studies show that the method of modeling score distribution is beneficial to various applications, such as outlier detection algorithms (Gao & Tan, 2006), search engines (Manmatha, Feng, & Rath, 2001), information filtering (Zhang & Callan, 2001), distributed information retrieval (Baumgarten, 1999), video retrieval (Wilkins, Ferguson, & Smeaton, 2006), kernel type selection for image retrieval (Doloc-Mihu & Raghavan, 2006), and biometry (Ulery, Fellner, Hallinan, Hicklin, & Watson, 2006).

The advantage of the score distribution method is that it uses the statistical properties of the scores, and not their values, and therefore, the obtained estimation may generalize better to not seen items than an estimation obtained by using the score values (Arampatzis, Beney, Koster, & van der Weide, 2000). In this chapter, we present the score distribution modeling approach, and then, we briefly survey theoretical and empirical studies on the distribution models, followed by several of its applications.

BACKGROUND

The primary goal of information retrieval is to retrieve all the documents which are relevant to a user query, while retrieving as few non-relevant documents as possible (Baeza-Yates & Ribeiro-Neto, 1999). This is achieved by ranking the list of documents according to their relevance to the user's query. Since relevance is a subjective attribute, depending on the user's perception of the closeness between the user submitted query and the real query from her or his mind, building a better way to retrieve data is a challenge that needs to be addressed in a retrieval system.

In other words, a retrieval system aims at building the request (query) that best represents the user's information need. This optimal request is defined by using an explicit data-request matching (Rocchio, 1971) that should produce a ranking in which all relevant data are ranked higher than the non-relevant data. For the matching process, a retrieval system uses a retrieval function, which associates each data-query pair with a real number or score (the retrieval status value). Then, the retrieval system uses these scores to rank the list of data.

However, researchers (Swets, 1963; Arampatzis, Beney, Koster, & van der Weide, 2000; Manmatha, Feng, & Rath, 2001) raised the question of whether or not the statistical properties of these scores, displayed by the shape of their distribution, for a given query, can be used to model the data space or the retrieval process. As a result, they proposed and empirically investigated several models of the score distributions as solutions to improve the effectiveness and efficiency of the retrieval systems. The next section introduces the score distribution method.

MAIN FOCUS

The Score Distribution Method

The probability ranking principle (Robertson, 1977) states that a search system should rank output in order of probability of relevance. That is, the higher the score value of the document, the more relevant to the query is considered the document to be. In the binary relevance case, which is the case we are interested in, the ideal retrieval system associates scores to the relevant and non-relevant data such that the two groups are well

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

separated, and relevant data have higher scores than the non-relevant data. In practice, retrieval systems are not capable to completely separate the relevant from the non-relevant data, and therefore, there are non-relevant data with higher score values than those of some relevant data.

The score distribution method tries to find a good way to separate these two groups of data by using statistical properties of their scores. The method assumes that the relevant and non-relevant data form two separate groups, with each group being characterized by its own characteristics different from the other group. For each group, the method plots the corresponding score values within the group, and then, tries to find the shape of the curve generated by these scores. In fact, this curve is approximated with a distribution usually chosen via experimental results (the best fit from a set of known distributions, such as normal, exponential, Poisson, gamma, beta, Pareto). Once the two distributions are known (or modeled), they are used to improve the search system.

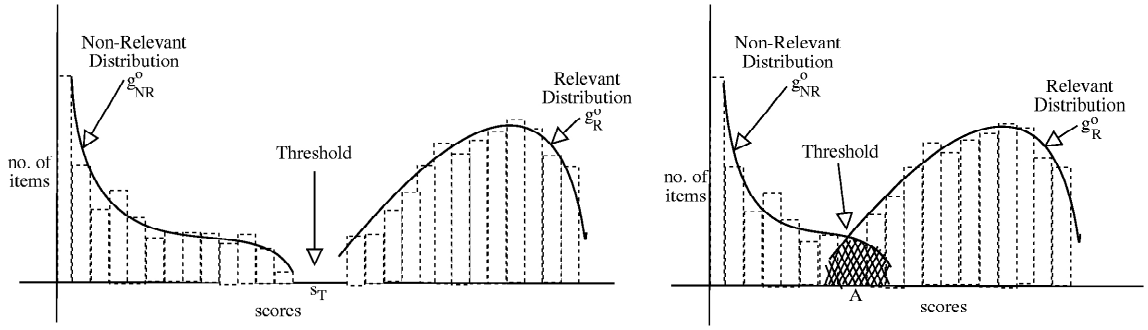
Figure 1 illustrates the score distribution method, (a) in the ideal case, when the relevant and non-relevant data are well separated by the retrieval system, and (b) in a real case, when there are non-relevant data with score values higher than those of relevant data. The scores of non-relevant data are grouped toward the left side of the plot, and the scores of relevant data are grouped toward the right side of the plot. A curve shows the shape of the score distribution of each group (of relevant and non-relevant data, respectively). Note that, in this figure, the two curves (given as densities $g^{O}_{R}(s)$ and $g^{O}_{NR}(s)$) do not display any particular distribution; they represent the curves of some arbitrary distributions. Basically, the score distribution method consists in choosing the best possible shapes of the distributions of the two groups. Then, any relevant (non-relevant) data is assumed to follow its chosen relevant (non-relevant) distribution.

Ideally, the two distribution curves do not meet (Figure 1 (a)), but in reality, the two curves meet at some point. However, as shown in Figure 1 (b), there is a common region between the two score distribution curves (named A). This area is of most interest for researchers; it includes relevant data with score values very close (lower or not) to the score values of non-relevant data. Therefore, by finding a way to minimize it, one finds a way to approximately separate the two data. Another solution is to find a threshold that separates optimally the relevant data from non-relevant ones.

The advantage of the score distribution method is that it uses the statistical properties of the scores (the shape of their distribution) and not their values, which conducts to an estimation of the threshold or the area A (Figure 1 (b)) that may generalize better to not seen data than an estimation method, which uses the score values (Arampatzis, Beney, Koster, & van der Weide, 2000).

We presented the method in the case that the entire data from collection is used. However, for efficiency reason, in practice, researchers prefer to return to user only the top most relevant N data. In this case, as Zhang and Callan (2001) noticed, the method is

Figure 1. Score distributions for relevant and non-relevant data



(a) Ideal case, at which a retrieval system aims, with a clear separation between the relevant and non-relevant data.

(b) Real case, which shows a common region for scores of the relevant and non-relevant data.

biased, especially for low scoring items, which do not occur between these top N chosen items. However, for high scoring data, the model offers a relatively good estimation (Zhang & Callan, 2001; Manmatha, Feng, & Rath, 2001).

Models of Score Distributions

Since introduced by Swets (1963), researchers used various combinations of the two distributions. Some chose the same type of distribution for both groups of data, whereas others argued that these should have different shapes. For example, Swets (1963) proposed two normal distributions of equal variance and later, two unequal variance normals or two exponentials (Swets, 1969); Bookstein (1977) used two Poisson distributions; Baumgarten (1999) used two gamma distributions. From the proposed models that use different distributions, the Gaussian-exponential model, which uses a normal for relevant data and an exponential for non-relevant data, is the most used model (Arampatzis & van Hameren, 2001; Manmatha, Feng & Rath, 2001; Zhang, & Callan, 2001; Collins-Thompson, Ogilvie, Zhang & Callan, 2003; Gao & Tan, 2006; Wilkins, Ferguson & Smeaton, 2006; Doloc-Mihu & Raghavan, 2006).

As shown by these examples, researchers investigated different specific distributions, such as normal, exponential, gamma, Poisson, but, to date, there is no agreement on either one of them as being the best distribution that models the scores or either relevant or not-relevant data. As Robertson (2007) noted recently, "clearly a strong argument for choosing any particular combination of distributions is that it gives a good fit to some set of empirical data" (p. 40). However, researchers addressed this issue in two ways. Some researchers base their models on the empirical evidence, while others try to find theoretical evidence. In the following, we briefly present such recent work on score distributions.

Theoretical Advances on Score Distribution Models

Rijsbergen (1979) observed that for search engines like SMART there is no evidence that the two score distributions should have similar shapes or that they follow Gaussian distributions as proposed by Swets (1963). Recently, some researchers try to find theoretical evidence to this observation. For example, Madigan, Vardi & Weissman (2006) presented an interesting mathematical analysis on the different combinations of distributions. They applied the extreme value theory (Resnick, 1987) to study why early precision increases as collection size grows. Their analysis showed that the asymptotic behavior of two retrieval measures (of effectiveness), P@K, the proportion of the top K documents that are relevant, and C@K, the number of non-relevant documents amongst the top K relevant documents, depends on the score distributions and on the relative proportion between relevant and non-relevant documents in the collection. The results contradict the normal-exponential model of Manmatha et al. (2001), and sustain Swets (1963) model with the remark that different choices (like, exponential-exponential, Pareto-Pareto, Beta-Beta) can result in early precision approaching zero or one or a constant as the number of ranked documents increases.

Robertson (2007) proposes the convexity hypothesis, which is a generalization of the hypothesis of the inverse recall-precision relationship and states that "for all good systems, the recall-fallout curve is convex" (p. 43). This hypothesis can be formulated as a condition on the probability of relevance of a document at an exact score: the higher the score, the higher the probability of relevance. The author proves that models like exponential-exponential (Swets, 1969), normal-normal with equal variances (Swets, 1963), Poisson-Poisson (Bookstein, 1977), gamma-gamma (Baumgarten, 1999) for certain settings of the parameters b and c, hold the convexity condition, and that the normal-normal model with different variances and the normal-exponential model (Manmatha, Feng & Rath, 2001) violate the condition. In conclusion, this theoretical result shows that the distribution models, which do not hold the convexity hypothesis, do not provide general solutions, but they are just reasonable approximations to the real distributions.

Empirical Studies on Score Distribution Models

Manmatha, Feng, & Rath (2001) show empirically that Gaussian-exponential models can fit approximately well the score distributions of the relevant and non-relevant documents corresponding to a given query. Moreover, they propose a model in which these score distributions are used to calculate the posterior prob-

abilities of relevance given the score via Bayes's rule. Experimental results on TREC-3, TREC-4, and TREC-6 data show that this method works for both probabilistic search engines, like INQUERY, and vector space search engines, like SMART, but it offers only an empirical approximation to the real distributions. In the same study, the authors also show that when relevance information is not available, these distributions can be recovered via the expectation-maximization algorithm by fitting a mixture model consisting of a Gaussian and an exponential function.

Applications of Score Distribution Method

As noted by Manmatha et al. (2001), once known, the score distributions can be used to map the scores of a search engine to probabilities. Score distributions can be beneficial to several tasks. A first example concerns the combination of the outputs of different search engines operating on one or more databases in different languages or not. This combination can be performed for example, by averaging the probabilities, or by using the probabilities to select the best engine for each query (Manmatha, Feng, & Rath, 2001).

Another example deals with the task of filtering thresholds. Here, Arampatzis and van Hameren (2001) proposed a score-distributional threshold optimization method for adaptive binary classification tasks. The method was tested on the TREC-9 Filtering Track and obtained the best results when using a Gaussian to model the distribution scores of the relevant documents, and an exponential for the distribution scores of the non-relevant documents. Zhang and Callan (2001) propose an algorithm that addresses the bias aspect of training data in information filtering, which happens because relevant information is not available for documents with scores below the threshold. Based on the Maximum Likelihood Principle, this algorithm estimates the parameters of the two score distributions (Gaussian-exponential model) and the ratios of the relevant and the non-relevant documents. The authors report significant improvement on the TREC-9 Filtering Track.

Baumgarten (1999) proposed a probabilistic solution based on a gamma-gamma model to select and fuse information from document subcollections over a distributed document collection. The model integrates acceptable non-heuristic solutions to the selection and fusion issues in a distributed environment, and shows encouraging experimental results that outperforms its non-distributed counterpart.

Gao and Tan (2006) proposed two approaches to convert output scores from outlier detection algorithms into well-calibrated probability estimates. Their second approach is similar to the one proposed by Manmatha et al. (2001) for search engines; it models the score distributions of outliers as a mixture of a Gaussian and an exponential probability function and calculates the posterior probabilities via Bayes's rule. The reported results show that the method helps in improving the selection of a more appropriate outlier threshold, and in improving the effectiveness of an outlier detection ensemble. Also, as in the case of search engines, the missing labels of outliers can be considered as hidden variables that can be learnt via the expectation-maximization algorithm together with the distribution model parameters.

Recently, the score distribution method was applied to multimedia data. For example, Doloc-Mihu and Raghavan (2006) used score distribution models to address the problem of automatically selecting the kernel type (Cristianini & Shawe-Taylor, 2000; Chapelle, Haffner, & Vapnik, 1999) for a given query in image retrieval. The authors empirically observed a correlation between the different kernel types and the different shapes of the score distributions. The proposed method selects the kernel type for which the surface of the intersection area (A) between the two distributions (see Figure 1 (b)) is minimal. The best retrieval results were obtained for the Gaussian-exponential model of the relevant and non-relevant images represented by color histograms in RGB color space. Further, this model gave also the best fit for fused multi-modal data (Doloc-Mihu & Raghavan, 2007).

Wilkins, Ferguson and Smeaton (2006) proposed a model based on score distributions to automatically generate the weights needed for multi-modal data fusion in video retrieval. Their model was formulated based on empirical observations, and compares features according to the score distribution of the scores of documents returned by them on a per query basis. Reported results on TRECVid 2004 and 2005 collections demonstrate the applicability of the model.

Finally, we mention the study performed recently by Ulery et al. (2006) of score-level fusion techniques involving different distribution models for biometric data. The authors used fingerprint and face data to evalu-

ate the effectiveness of eight score fusion techniques. The study concluded that fusing scores is effective, but it depends on a series of factors such as the ability to model score distributions accurately.

FUTURE TRENDS

Future research should focus on developing a well-founded theoretical model for choosing the score distributions that describes the correlations, empirically observed, between score distributions and relevance. Further, such a model should be tested on real-world data to investigate its entire potential.

Different Data Mining tasks, such as mining Web information, mining multimedia, and biomedical data, and information retrieval tasks, such as multi-lingual retrieval, and relevance feedback may benefit from modeling score distributions. Other potential applications such as feature selection and fusion for image, video and sound retrieval need to be considered for multimedia engines.

CONCLUSION

In this chapter, we presented the score distribution modeling approach, which offers elegant solutions to improve the effectiveness and efficiency of different components of search systems. We surveyed several such models used for various tasks, such as ensemble outlier detection, finding thresholds for document filtering, combining the outputs of different search engines, selecting the kernel type in an Image Retrieval System, and fusion of multimodal data in video retrieval and biometry. These applications demonstrate, mostly through empirical testing, the potential of the score distribution models to real world search systems.

REFERENCES

Arampatzis, A., Beney, J., Koster, C., & van der Weide, T.P. (2000). Incrementally, Half-Life, and Threshold Optimization for Adaptive Document Filtering. *IEEE Transactions on Knowledge and Data Engineering*, 15(6), 1460-1471.

Arampatzis, A., & van Hameren, A. (2001). The Score-Distributional Threshold Optimization for Adaptive Binary Classification Tasks. In W.B. Croft, D.J. Harper, D.H. Kraft, J. Zobel (Eds.), *Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information retrieval* (pp.285-293). New York: ACM Press.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information retrieval*, New York: Addison-Wesley.

Baumgarten, C. (1999). A Probabilistic Solution to the Selection and Fusion Problem in Distributed Information retrieval. In M. Hearst, F. Gey, R. Tong (Eds.), *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information retrieval, Univ. of California at Berkeley, USA,* (pp.1-8). New York: ACM Press.

Bookstein, A. (1977). When the most 'pertinent' document should not be retrieved – an analysis of the Swets model. *Information Processing and Management*, 13(6), 377-383.

Chapelle, O., Haffner, P., & Vapnik, V. (1999). SVMs for Histogram-Based Image Classification. *IEEE Transactions on Neural Networks*, 10(5), 1055-1064.

Collins-Thompson, K., Ogilvie, P., Zhang, Y., & Callan, J. (2003). Information filtering, novelty detection and named page finding. In E.M. Voorhees, L.P. Buckland (Eds.), *The 11th Text retrieval Conference, TREC 2002, NIST Special Publication 500-251 (NIST 2003)*, *Gaithersburg, Maryland, USA,* (pp. 107-118).

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to Support Vector Machines and other kernel-based learning methods*, New York: Cambridge University Press.

Del Bimbo, A. (2001). *Visual Information retrieval,* San Francisco, CA: Morgan Kaufmann Publishers, Inc.

Doloc-Mihu, A. (2007). *Adaptive Image Retrieval: Similarity Modeling, Learning, Fusion, and Visualization*, Ph.D. dissertation. Louisiana, USA: University of Louisiana at Lafayette.

Doloc-Mihu, A., & Raghavan, V. V. (2006). Score Distribution Approach to Automatic Kernel Selection for Image Retrieval Systems. In F. Esposito, Z.W. Ras, D. Malerba, G. Semeraro (Eds.), *Proceedings of the*

16th International Symposium on Methodologies for Intelligent Systems (ISMIS 2006) Bari, Italy: LNAI, Vol.4203, *Foundations of Intelligent Systems* (pp. 443-452). Berlin: Springer-Verlag.

Doloc-Mihu, A., & Raghavan, V. V. (2007). Fusion and Kernel Type Selection in Adaptive Image Retrieval. In B.V. Dasarathy (Ed.), *Multisensor, Multisource Information Fusion, SPIE Defense and Security Symposium, (DSS 2007), Vol.6571, Orlando, Florida, USA* (pp. 75-90). Washington: SPIE Press.

Gao, J., & Tan, P.-N. (2006). Converting Output Scores from Outlier Detection Algorithms into Probability Estimates. In B. Werner (Ed.), *Proceedings of the 6th International Conference on Data Mining (ICDM 2006), Hong Kong,* (pp. 212-221). Los Alamitos: IEEE Computer Society.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Technique,*. San Francisco, CA: Morgan Kaufmann Publishers, Inc.

Madigan, D., Vardi, Y., & Weissman, I. (2006). Extreme Value Theory Applied to Document Retrieval from Large Collections. *Information retrieval,* 9(3), 273-294.

Manmatha, R., Feng, F., & Rath, T. (2001). Using Models of Score Distributions in Information retrieval. In W.B. Croft, D.J. Harper, D.H. Kraft, J. Zobel (Eds.), *Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information retrieval, New Orleans, USA,* (pp.267-275). New York: ACM Press.

Resnick, S.I. (1987). *Extreme Values, Regular Variation and Point Processes,* New York: Springer-Verlag.

Rijsbergen, C.J. (1979). *Information retrieval,* Butterworths, London, 2nd Ed.. Retrieved December 9, 2007, from http://www.dcs.gla.ac.uk/Keith/Preface.html.

Robertson, S.E. (1977). The probability ranking principle in information retrieval. *Journal of Documentation,* 33(4), 294-304.

Robertson, S. (2007). On score distributions and relevance. In G. Amati, C. Carpineto, G. Romano (Eds.), *29th European Conference on Information retrieval, ECIR 2007, Rome, Italy, LNCS, vol. 4425* (pp. 40-51). Berlin: Springer-Verlag.

Rocchio, J.J. (1971). Relevance feedback in Information retrieval. In Gerard Salton (Ed.), *The SMART Retrieval System - Experiments in Automatic Document Processing* (pp. 313-323). Englewood Cliffs, New Jersey, USA: Prentice-Hall, Inc.

Swets, J. A. (1963). Information retrieval Systems. *Science*, 141(3577), 245-250.

Swets, J. A. (1969). Effectiveness of Information retrieval Methods. *American Documentation*, 20, 72-89.

Ulery, B., Fellner, W., Hallinan, P., Hicklin, A., & Watson, C. (2006). Studies of Biometric Fusion. *NISTIR*, *7346*. Retrieved October 10, 2007, from http://www.itl.nist.gov/

Zhang, Y., & Callan, J. (2001). Maximum Likelihood Estimation for Filtering Thresholds. In W.B. Croft, D.J. Harper, D.H. Kraft, J. Zobel (Eds.), *Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information retrieval, New Orleans, USA,* (pp.294-302). New York: ACM Press.

Wilkins, P., Ferguson, P., & Smeaton, A. F. (2006). Using Score Distributions for Query-time Fusion in Multimedia Retrieval. In J.Z. Wang, N. Boujemaa (Eds.), *Proceedings of the 8th ACM International Workshop on Multimedia Information retrieval, MIR 2006, Santa Barbara, California, USA,* (pp. 51-60). New York: ACM Press.

Wikipedia (2007a). Information retrieval. Retrieved December 8, 2007, from http://en.wikipedia.org/wiki/Information_retrieval.

Wikipedia (2007b). Kernel Methods. Retrieved December 8, 2007, from http://en.wikipedia.org/wiki/Kernel_Method.

KEY TERMS

Fusion: Process of combining two distinct things. **Data fusion** and **Information Fusion** used in Data Mining are generally defined as the set of techniques that combine/merge data/information from multiple sources.

Information Filtering: Process of monitoring information in order to present to the user information items the user is interested in.

Information Retrieval: Science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within databases, whether relational stand-alone databases or hypertextually-networked databases such as the World Wide Web (Wikipedia, 2007a). Similarly, Image Retrieval is the science of searching and retrieving images from a large database of digital images (del Bimbo, 2001; Doloc-Mihu, 2007).

Kernel Methods: Class of algorithms for pattern analysis, which use kernel functions that allow to operate in the feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the corresponding features of all pairs of data in the feature space (Wikipedia, 2007b).

Kernel Type Selection: Process of selecting the form of the kernel function.

Expectation-Maximization Algorithm: An iterative algorithm used for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved hidden variables.

Outlier: Data objects from a database, which do not comply with the general behavior or model of the data (Han and Kamber, 2006). **Outlier detection and analysis** is an important data mining task, named **outlier mining**, with applications, for example, in fraud detection.

Score Distribution Model: Model associated with a particular combination of the distributions of the score values of relevant and non-relevant data. There are several models proposed so far, such as exponential-exponential, normal-normal, Gaussian-exponential, gamma-gamma, and so on. However, a good reason for choosing any particular model is based on how good the distributions fit a set of empirical data.

TREC: Text Retrieval Conference that focuses on different information retrieval research areas, or tracks, and provides the infrastructure necessary for large-scale evaluation of text retrieval methodologies.

TRECVID: Is conference separate from the TREC conference that focuses on a list of different information retrieval (IR) research areas in content based retrieval of video.