WILEY InterScience® DISCOVER SOMETHING GREAT

BRIEF COMMUNICATION

On Relevance Distributions

Wolfgang G. Stock
Department of Information Science, Heinrich-Heine-University Düsseldorf, Universitätsstraße 1, D-40225 Düsseldorf, Germany. E-mail: stocknmw@aol.com

There are at least three possible ways that documents are distributed by relevance: informetric (power law), inverse logistic, and dichotomous. The nature of the type of distribution has implications for the construction of relevance ranking algorithms for search engines, for automated (blind) relevance feedback, for user behavior when using Web search engines, for combining of outputs of search engines for metasearch, for topic detection and tracking, and for the methodology of evaluation of information retrieval systems.

Introduction

Relevance is a key concept in information science—and "a key headache," Tefko Saracevic (1999, p. 1058) adds. For Saracevic *relevance* is "the attribute or criterion reflecting the effectiveness of exchange of information between people (i.e., users) and information retrieval (IR) systems in communication contacts based on valuation by people" (1999, p. 1059). A *relevance distribution* is a distribution of documents (e.g., full texts or bibliographic records) sorted by relevance judgments. These relevance judgments are both user (or expert) judgments and weighting attributes (e.g., weighted keywords in a record or the number of citations of an academic document). There is extensive literature about relevance in information science (see, e.g., Greisdorf, 2000; Journal of the American Society for Information Science, 1994; Mizzaro, 1997; Saracevic, 1975; Schamber, 1994), and there are some remarks on relevance distributions as well (e.g., Spink & Greisdorf, 2001, p. 163; Greisdorf & Spink, 2001, p. 845; Lavrenko, 2004, pp. 27–31).

The nature of the "right" relevance distribution has implications. (1) The notion of relevance is crucial for all those processes called *relevance ranking*. If we do not know how documents are generally distributed by relevance, it is not possible to create a reasonable algorithm of relevance ranking. (2) It is also crucial for search engine users to make the decision about many of the top ranked documents they should read. (3) If a retrieval system works with automated, blind relevance feedback, it is important to know how many documents of the first ranking procedure are to be considered in the second step and are to be analyzed. (4) If we want to model score distributions for combining the outputs of search engines for metasearch, we need to know the nature of these distributions. (5) For topic detection and tracking, i.e., finding similar documents (e.g., news) on the same topic, relevance models are useful. (6) Evaluation exercises in information retrieval (e.g., Text Retrieval Conferences [TReC]) usually work with a dichotomous view of relevance. So it is possible that an information retrieval system whose scores approximate the "right" relevance distribution (whichever it is) better than other systems does not receive a better evaluation assessment.

Relevance Distributions

There are (at least) three possible kinds of relevance distributions, (1) a dichotomous view, (2) an informetric view, and (3) an inverse logistic view.

In Figure 1 the *Y* axis represents the degree of relevance (between 1 und 0) and the *X* axis the documents of rank 1, 2, 3, and so on. In our example we see a hypothetical distribution of 20 documents. In fact, when using Google or Yahoo! a much higher number of hits can usually be expected. So our rank 10 may be rank 1,000 or 10,000 or even higher in real search results lists.

The Dichotomous View of Relevance Distribution

In the dichotomous view, the users prefer a binary decision (Janes, 1991). A document is either relevant or not relevant to a specific query. In the dichotomous view, some documents are relevant (with a 1), all others are irrelevant (with a 0). All Boolean information retrieval systems implicitly work with this assumption. Of course, no relevance ranking is possible

Received February 17, 2004; revised April 13, 2005; accepted April 28, 2005

 ● Published online 6 April 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20359

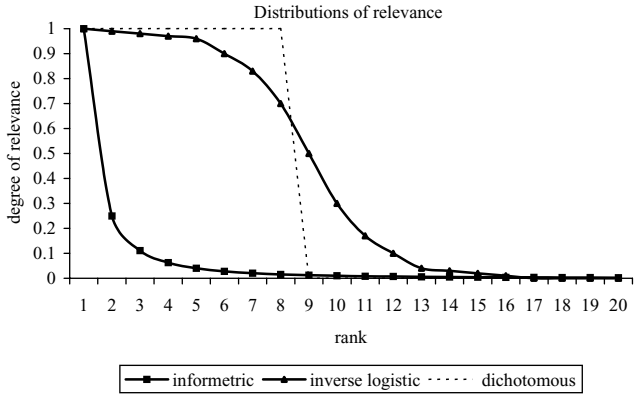

FIG. 1. Possible relevance distributions: informetric, inverse logistic, dichotomous.

here; other criteria for sorting (such as date) are usual. Most of the professionals in information research prefer this dichotomous view. Although, for example, some information service providers (DIALOG, Lexis-Nexis, or Westlaw) have introduced relevance ranking functionality (see, e.g., Tenopir & Cahn, 1994), most information professionals do not use it in their practical work. For Janes (1993, p. 113) the dichotomous view could be an artifact because when evaluating retrieval results people make these yes-no decisions because they have been asked to do so as part of a research study.

Maron and Kuhns (1960) argued that relevance is not a 1/0 decision and, on the basis of this assumption, introduced their probabilistic model of information retrieval. In this tradition, Robertson (1977) put particular emphasis on the probabilistic character of relevance. Spink and Greisdorf (2001) found hints for "partial relevance" in their user studies. The probabilistic IR model, the factual perception of relevance by users, and the very practical reason of designing relevance ranking in Web search engines led to the opinion that the dichotomous view of relevance distributions is not very helpful.

The Informetric View of Relevance Distribution

In the tradition of the laws of Zipf, Bradford, and Lotka there is a general law of informetric distributions that follows the formula

$$f(x) = \frac{C}{x^a}$$

(Egghe & Rousseau, 1990, p. 293; Stock, 2000, p. 130), where C is a constant, x is the rank, and a is a value ranging normally from about 1 to about 2 (in Figure 1, it is 2). For large sets this power law proves to be correct in many examples (see, e.g., Saracevic, 1975, pp. 329–331). The maximal value in relevance distributions is 1, so here C is 1 as well. The law tells us that if the document of rank 1 has a relevance of 1, then the second ranked document must have a relevance of 0.5 (with $a = 1$) or of 0.25 (with $a = 2$), the third 0.33 ($a = 1$) or 0.11 ($a = 2$), and so on.

The Inverse Logistic View of Relevance Distribution

In a research project concerning the relevance of scientific documents in relation to given topics (Stock, 1981) we found out that about 14.5% of all documents in the hit list had a weight of 100 (i.e., a relevance of 1), very few items had values between 99 and 31, and about 66% of the documents had values between 30 and >0. The relevance judgment was made by the indexer. This example may stand for another kind of relevance distribution. Data by Spink and Greisdorf (2001, p. 165) lead in the same direction. Here we see several documents having a 1 or a close to 1 value on the first ranks, then in the middle there are relatively few items, followed by a large set of documents with low relevance weights. If one looks at the function one can see a graph similar to the logistic function but here it is mirrored. So we speak of an inverse logistic function. The mathematical expression of this function is

$$f(x) = e^{-C'(x-1)^b}$$

where e is the Euler number, x is the rank, C' is a constant, and the exponent b is approximately 3. In our example b is 3 and C' is 0.0015.

The size of the hit set. There is evidence that both views, the informetric and the inverse logistic, work in certain hit sets with many documents. In smaller sets these laws are not valid in all cases. The smaller the set the smaller is the probability that there will be an informetric or an inverse logistic distribution. In small hit sets it is very probable that there is no regularity of relevance distribution. If a retrieval system output consists of three hits, one discussing the searched topic on half a page in an article of 10 pages and two mentioning the topic only in short footnotes, then you cannot decide whether this follows an informetric or an inverse logistic function (in Figure 1 our three documents could be positioned in the area of our ranks 4 to 6 in the informetric function or at ranks 13 to 15 in the inverse logistic function) or whether there is no regularity at all.

Implications

Implications for Blind Relevance Feedback

In the probabilistic retrieval model one needs relevance information to calculate the relevance of a document under a specific query. It is possible to present to users a list of hits generated by another retrieval model (say, Google-like or Kleinberg-like link topology or Salton's vector space model) in the first round and let the user decide whether a document is relevant for him or not. Now we have relevance information and can analyze the words in the texts that are mentioned either as relevant or not relevant. There is another possibility for obtaining relevance information. Croft and Harper (1979) work without user judgments. In their *pseudorelevance feedback model* there is a basic assumption: the top-ranked

DOI: 10.1002/asi

documents are relevant and lower-ranked items are not. Here the relevance feedback procedure works "blindly" and automatically. But which top-ranked items are to be considered? If there is an informetric distribution, the system should analyze only a few documents, but if there is an inverse logistic distribution (and a large hit set), the retrieval system should analyze all documents at 1 or close to 1. So a good strategy under the informetric view is a bad decision under the inverse logistic view.

Implications for Topic Detection and Tracking

Topic detection and tracking (TDT) includes the segmentation of a news stream into stories about a single topic, the detection of a new topic, and the monitoring of the news stream for additional stories on a known topic. For this last aspect of tracking one can work with the Croft-Harper pseudorelevance feedback model. "When we estimate a Relevance Model for some story, we mix together neighboring document models" (Lavrenko et al., 2002). How many documents are neighbors? It is the same problem that we just mentioned concerning the blind relevance feedback.

Implications for User Behavior

Using AltaVista data, Silverstein, Henzinger, Marais, and Moricz (1998) observed that 85.2% of all users prefer to look only at the first result screen (with 10 records). Spink, Wolfram, Jansen, and Saracevic (2001) observed slightly different user behavior: about 28% of all users of the search engine Excite pay attention only to the first 10 hits, 19% to the first 20, and 12% to the first 30 hits. There is one main finding in all research projects of user behavior with Web search engines: users prefer just to click on the first 10 to 30 links. If the inverse logistic view were true (and if there is a big hit set), this user behavior would be wrong because the user could not see documents with top relevance. But if the informetric view is true (and if the query was perfect and the relevance ranking algorithm works ideally), the observed user behavior is strongly rational. Under the informetric view (even with very big hit sets) you can expect the best hits at the top of the distribution—and looking only at the first 30 entries is a time-saving action that is sufficient for most information needs of end users.

Implications for Data Fusion in Metasearch

Algorithmic search engines rank hits by machine relevance "judgments," so that the most relevant item is ranked first, the second most relevant is ranked second, and so on. For every document a specific score is calculated, but different search engines work with different algorithms. If we want to merge hit lists from different search engines, we need a function to put a document ranked *x*th by search engine 1 into a combined list of all documents found by all search engines. According to Manmatha and Sever (2002) and Manmatha, Rath, and Feng (2001) it is possible to map the scores of a search engine to probabilities and sort the merged list by the calculated probability. Another possibility is to work with ranks instead of scores or to work with both ranks and scores (Hsu & Taksa, 2005). A necessary condition for this mapping is a model of score distribution. And this model depends on the "right" model of relevance distribution.

Implications for Evaluation Tasks

The most common measures of recall and precision of retrieval systems are based on the dichotomous view of relevance and (in the past) on a binary view of retrieval results (either a document is retrieved or it is not). "These assumptions can, and need to, be questioned: relevance might be not binary, and IRS (Information Retrieval Systems) usually rank the retrieved documents and, sometimes, show their weight" (Della Mea, Di Gaspero, & Mizarro, 2004, p. 30). The second problem can be solved by the introduction of a cut-off value: evaluators consider only the first x documents of a ranked results list und calculate both the average precision of the top x hits and the rank-specific precision values (precision@1, precision@2, and so on). In a recent study Lo Grasso and Wahlig (2005) work with $x = 20$. But the first problem is still open. So it is not possible inside, e.g., the framework of TReC to study whether a system approximates the "right" relevance distribution or not. Perhaps there is a solution in the average distance measure (*ADM*) of Della Mea and Mizzaro (2004, p. 533), who introduce a user relevance score (*URS*) that measures the relevance of a document with respect to an information need of a certain person. *URS* is an expression of subjective relevance judgments, which assumes various values between 0 and 1. The system relevance score (*SRS*; also in the [0, 1] range) is the objective relevance "judgment" of a given IR system. The new retrieval effectiveness measure *ADM* is the average distance (or difference) between *URS* and *SRS*. With the help of *URS*, *SRS*, and *ADM* it will be possible to begin evaluation research concerning the kinds of relevance distribution as well. (There is a practical problem if search engines do not show their relevance scores.)

What Now?

The methodology of measuring effectiveness of IR systems by using the binary view (e.g., by TReC) can be very problematic, for it does not permit one to observe and to analyze the specific natures of nondichotomous relevance distributions and the behavior of IR systems to adapt to these natures.

There are completely different implications, e.g., for rational user behavior using search engines, for working with blind relevance feedback and for constructing ranking algorithms for search engines if the informetric or the inverse logistic view of relevance distribution is right (the dichotomous view is not applicable on the Web).

Is only one view correct and the other false? Perhaps both views are false, and there is no general regularity in

DOI: 10.1002/asi

relevance distributions at all. Or are both views true for different kinds of relevance? And what role does the size of the hit set play? Our "key headache" leads not only to the known question "How many relevances in information retrieval?" (Mizzaro, 1998), but also to "How many relevance distributions?"

Acknowledgments

I would like to thank Jasmin Schmitz for checking my English, Arnold Kobalz for mathematical hints, and the anonymous referee for useful suggestions.

References

Croft, W.B., & Harper, D.J. (1979). Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35, 285–295.

Della Mea, V., Di Gaspero, L., & Mizzaro, S. (2004). Evaluating ADM in a four-level relevance scale document set from NTCIR. In NTCIR Workshop 4 Meeting. Working Notes of the Fourth NTCIR Workshop Meeting (Vol. 2, Suppl., pp. 30–38). Tokyo: National Institute of Informatics.

Della Mea, V., & Mizzaro, S. (2004). Measuring retrieval effectiveness: A new proposal and a first experimental validation. Journal of the American Society for Information Science and Technology, 55, 530–543.

Egghe, L., & Rousseau, R. (1990). Introduction to informetrics. Amsterdam: Elsevier.

Greisdorf, H. (2000). Relevance: An interdisciplinary and information science perspective. Informing Science, 3(2), 67–71.

Greisdorf, H., & Spink, A. (2001). Median measure: An approach to IR systems evaluation. Information Processing & Management, 37, 843–857.

Hsu, D.F., & Taksa, I. (2005). Comparing rank and score combination methods for data fusion in information retrieval. Information Retrieval, 8, 449–480.

Janes, J.W. (1991). The binary nature of dichotomous relevance judgments: A study of users' perceptions. Journal of the American Society for Information Science, 42, 745–756.

Janes, J.W. (1993). On the distribution of relevance judgments. In Proceedings of ASIS 1993 (pp. 104–114). Medford, NJ: Learned Information.

Journal of the American Society for Information Science. (1994). Relevance research [Special issue]. Journal of the American Society for Information Science, 45(3).

Lavrenko, V. (2004). A generative theory of relevance. Unpublished dissertation, Computer Science, University of Massachusetts Amherst.

Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., & Thomas, S. (2002). Relevance models for topic detection and tracking. In Proceedings of the Human Language Technology Conference, San Diego 2002. Retrieved from ciir.cs.umass.edu/pubfiles/ir-243.pdf

Lo Grasso, L., & Wahlig, H. (2005). Google und seine Suchparameter: Eine Top 20-Precision Analyse anhand repräsentativ ausgewählter Anfragen. Information—Wissenschaft und Praxis, 56, 77–86.

Manmatha, R., Rath, T., & Feng, F. (2001). Modeling score distributions for combining the outputs of search engines. In W.B. Croft (Ed.), Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 267–275). New York: ACM.

Manmatha, R., & Sever, H. (2002). A formal approach to score normalization for meta-search. In Proceedings of the Human Language Technology Conference, San Diego. Retrieved from www.baskent.edu.tr/~sever/ir-242.pdf

Maron, M.E., & Kuhns, J.L. (1960). On relevance, probabilistic indexing, and information retrieval. Journal of the Association for Computing Machinery, 7, 216–244.

Mizzaro, S. (1997). Relevance: The whole story. Journal of the American Society for Information Science, 48, 810–832.

Mizzaro, S. (1998). How many relevances in information retrieval? Interacting With Computers, 10, 303–320.

Robertson, S.E. (1977). The probabilistic character of relevance. Information Processing & Management, 13, 247–251.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26, 321–343.

Saracevic, T. (1999). Information science. Journal of the American Society for Information Science, 50, 1051–1063.

Schamber, L. (1994). Relevance and information behavior. Annual Review of Information Science and Technology, 29, 3–48.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1998). Analysis of a very large AltaVista query log (SRC Technical Note 1998-014). Palo Alto, CA: Digital Systems Research Center.

Spink, A., & Greisdorf, H. (2001). Regions and levels: Measuring and mapping users' relevance judgments. Journal of the American Society for Information Science, 52, 161–173.

Spink, A., Wolfram, D., Jansen, B.J., & Saracevic, T. (2001). Searching the Web: The public and their queries. Journal of the American Society for Information Science and Technology 52, 226–234.

Stock, W.G. (1981). Die Wichtigkeit wissenschaftlicher Dokumente relativ zu gegebenen Thematiken. Nachrichten für Dokumentation, 32, 162–164.

Stock, W.G. (2000). Informationswirtschaft: Management externen Wissens. München, Wien: Oldenbourg.

Tenopir, C., & Cahn, P. (1994). TARGET & FREESTYLE: DIALOG and Mead join the relevance ranks. Online, 18(3), 31–47.

DOI: 10.1002/asi