Towards Mathematical Foundations of Information Retrieval: Dependence of Website's Relevance on the Number of Occurrences of a Queried Word

László T. Koczy[1], Vladik Kreinovich[2], Yohans Mendoza[2],
Hung T. Nguyen[2] and Harry Schulte[4]

[1]Department of Telecommunication and Telematics,
Technical University of Budapest,
Budapest H-1521, Hungary, email koczy@boss.ttt.bme.hu

[2]Department of Computer Science and
[4]Multi-Media Learning and Tutoring Center
University of Texas at El Paso, El Paso, TX 79968, USA
emails {vladik,hschulte}@utep.edu

[3]Department of Mathematical Sciences, New Mexico State University,
Las Cruces, NM 880033, USA, email hunguyen@nmsu.edu

Abstract. *In response to a query, web search tools often return many websites which are not really relevant. One reason for this is that the queried word may have several meanings different to the one which the user has in mind. To eliminate these undesirable meanings, it is reasonable to look for occurrences not only of the queried word itself, but also for other words related to this particular meaning, and then select only the websites for which, based on this information, we are confident about their relevance. For this strategy to work, we must be able to estimate the degree of relevance d of a website based on the number of occurrences N of given word.*

In this paper, we describe the optimal model for the dependence $d(N)$.

Introduction. In response to a query, web search tools often return many websites which are not really relevant. For example, a query about the word "fuzzy" may return a webpage on which some person feels fuzzy about his relationship. One reason for such irrelevant websites is that the queried word may have several meanings different to the one which the user has in mind.

To eliminate these undesirable meanings, it is reasonable to look for occurrences not only of the queried word itself, but also for other words related to this particular meaning, and then select only the websites for which, based on this information, we are confident about their relevance. This strategy has been described and tested in [1].

For this strategy to work successfully, we must be able to estimate the degree (fuzzy value, subjective probability) of relevance d of a website based on the number of occurrences N of given word.

In this paper, we describe the optimal model for the dependence $d(N)$. Our justification for the resulting formula will use methods motivated by the neural network approach (see, e.g., [2]).

The values $d(N)$ depend on the pre-selection procedure. In principle, the values $d(N)$ can be determined as frequencies, from the statistical analysis of different queries. However, this statistics may be somewhat confusing because in reality, the frequency $d(N)$ depends on how we pre-select the webpages which are analyzed by the web search tool.

Some web search tools pride themselves on covering the largest possible amount of webpages; other web search tools pre-select the webpages based on the topic of the query and only look into those pages which a priori seem to be relevant. For example, if we ask about "fuzzy" having science and engineering applications in mind, then webpages about relationships would probably not be pre-selected and thus, would not appear in the search tool's answer to the query. The advantage of not having to look through millions of (most probably irrelevant) webpages is that more time is left for a more sophisticated analysis of each pre-selected website.

Depending on the pre-selection, we may have different dependencies $d(N)$. For example, if we do make a pre-selection, then the subjective probability that a page with a small value of N is relevant must be higher than without a pre-selection, because the very fact that the page has been pre-selected increases the chance that this page is relevant.

So, instead of looking for a *single* function $d(N)$, we should look for a *family* of functions which correspond to different pre-selections.

Relation between the functions $d(N)$ corresponding to different pre-selection procedures. How are different functions from this family related to each other? Pre-selection means, in effect, that we are moving from the original *unconditional* probability of relevance $d(N)$ to the *conditional* probability, under the condition that this particular page has been pre-selected. In statistics, the transformation from an unconditional probability $P_0(H_i)$ of a certain hypothesis H_i to its conditional probability $P(H_i|S)$ (under the condition S that the webpage was pre-selected) is described by the Bayes formula

$$P(H_i|S) = \frac{P(C|H_i)\cdot P_0(H_i)}{\sum\limits_j P(S|H_j)\cdot P_0(H_j)}.$$

In mathematical terms, the transformation from $d(N) = P_0(H_i)$ to $\widetilde{d}(N) = P(H_i|S)$ is *fractionally linear*, i.e., has the form $d(N) \to \widetilde{d}(N) = \varphi(d(N))$, where

$$\varphi(z) = \frac{k\cdot z + l}{m\cdot z + n}$$

for some real numbers k, l, m, and n.

Resulting description of the desired family of functions $d(N)$. So, instead of looking for a single function $d(N)$, we should look for a family of functions $\{\varphi(d(N))\}$, where $d(N)$ is a fixed function and $\varphi(z)$ are different fractionally linear transformations. In the following text, when we say "a family of functions", we will mean a family of this very type.

We can have many different optimality criteria. Among all such families, we want to choose the best one. In formalizing what "the best" means we follow the general idea outlined in [2]. The criteria to choose may be:

- approximation accuracy (i.e., accuracy with which these functions approximate the empirical data about the dependence of the probability of relevance d on the number of occurrences N),
- computational simplicity, etc.

Non-numeric criteria are possible. In mathematical optimization problems, *numeric* criteria are most frequently used, when to every family we assign some value expressing its performance, and choose a family for which this value is maximal. However, it is not necessary to restrict ourselves to such numeric criteria only. For example, if we have several different families that have the same approximation accuracy A, we can choose between them the one that has the minimal computational complexity C.

In this case, the actual criterion that we use to compare two families is not numeric, but more complicated: A family F_1 is better than the family F_2 if and only if either $A(F_1) > A(F_2)$, or $A(F_1) = A(F_2)$ and $C(F_1) < C(F_2)$.

A general description of optimality criteria. A criterion can be even more complicated than above. What a criterion *must* do is to allow, us for every pair of families (F_1, F_2), to tell:

- whether the first family is better with respect to this criterion (we'll denote it by $F_1 > F_2$),
- or the second is better $(F_1 < F_2)$,
- or these families have the same quality in the sense of this criterion (we'll denote it by $F_1 \sim F_2$).

Of course, it is necessary to demand that these choices be consistent, e.g., if $F_1 > F_2$ and $F_2 > F_3$ then $F_1 > F_3$.

A criterion must choose a unique optimal family. A natural demand is that this criterion must choose a *unique* optimal family (i.e., a family that is better with respect to this criterion than any other family). The reason for this demand is simple:

If a criterion does not choose any family at all, then it is of no use.

If several different families are "the best" according to this criterion, then we still have a problem to choose among those "best". Therefore, we need some additional criterion for that choice. For example, if several families turn out to have the same approximation accuracy, we can choose among them a family with minimal computational complexity. So what we actually do in this case is abandon that criterion for which there were several "best" families, and consider a new "composite" criterion instead: F_1 is better than F_2 according to this new criterion if either it was better according to the old criterion or according to the old criterion they had the same quality and F_1 is better than F_2 according to the additional criterion.

In other words, if a criterion does not allow us to choose a unique best family it means that this criterion is not ultimate; we have to modify it until we come to a final criterion that will have that property.

A criterion must be scale-invariant. There are several different ways of handling a query. For example, if we ask for the word "neural", we may take this query literally and only look for the occurrences of this very word "neural", or we can also take into consideration closely related words such as "neuron" and "neurons". On average, using several closely related words increases the number of occurrences. For example, if we use two words instead of one, then probably we do not have a double increase in the number of occurrences, but we will have a proportional increase with some coefficient λ between 1 and 2. Similarly, if we use three words instead of one, we may have, on average, a proportional increase with some coefficient λ between 1 and 3.

In general, whenever we had N occurrences, we will now have (on average) $\widetilde{N} = \lambda \cdot N$ occurrences. How will the dependence $d(N)$ change, i.e., what will be the new function $\widetilde{d}(\widetilde{N})$ describing the dependence of the probability of relation on the new number of occurrences $\widetilde{N}$? $\widetilde{N}$ new occurrences are equivalent to $N = \widetilde{N}/\lambda$ old occurrences, so the desired probability $\widetilde{d}(\widetilde{N})$ is equal to $\widetilde{d}(\widetilde{N}) = d(N/\lambda)$.

It is reasonable to require that the relative quality of two different families should not change if we simply change the way we count occurrences.

An alternative way to handle occurrences of similar words is to count them not as full occurrences, but as partial occurrences of the original word: e.g., we may count each occurrence of the word "neuron" as 0.8 of an occurrence of the word "neural". Since we are using fractional values, the resulting total number of occurrences N is not necessarily an integer. So, we must define $d(N)$ not only for integer values of N, but also for arbitrary real values N.

We arrive at the following definitions:

Definition 1. *By a relevance function, we mean a smooth monotonic function $d(N)$ defined for all real numbers $N \geq 0$ for which $d(0) = 0$ and $d(N) \to 1$ as $N \to \infty$.*

Definition 2. *By a family of functions we mean the set of functions that is obtained from a relevance function $d(N)$ by applying fractionally linear transformations.*

Definition 3. *A pair of relations $(<, \sim)$ is called consistent if it satisfies the following conditions: (1) if $F < G$ and $G < H$ then $F < H$; (2) $F \sim F$; (3) if $F \sim G$ then $G \sim F$; (4) if $F \sim G$ and $G \sim H$ then $F \sim H$; (5) if $F < G$ and $G \sim H$ then $F < H$; (6) if $F \sim G$ and $G < H$ then $F < H$; (7) if $F < G$ then $G < F$ or $G \sim F$ are impossible.*

Definition 4. *Assume a set $\mathcal{F}$ is given. Its elements will be called alternatives. By an optimality criterion we mean a consistent pair $(<, \sim)$ of relations on the set $\mathcal{F}$ of all alternatives. If $F > G$, we say that F is better than G; if $F \sim G$, we say that the alternatives F and G are equivalent with respect to this criterion. We say that an alternative F is optimal (or best) with respect to a criterion $(<, \sim)$ if for every other alternative G, either $F > G$ or $F \sim G$.*

Definition 5. *We say that a criterion is final if there exists an optimal alternative, and this optimal alternative is unique.*

In the present section we consider optimality criteria on the set $\mathcal{F}$ of all families.

Definition 6. *Let $\lambda > 0$. By the λ-rescaling $S_\lambda(p)$ of a function $d(N)$, we mean a function $\widetilde{d}(N) = d(N/\lambda)$. By the λ-rescaling $S_\lambda(F)$ of the family F, we mean the family of the functions that are obtained from $d \in F$ by λ-rescaling.*

Definition 7. *We say that an optimality criterion on $\mathcal{F}$ is scale-invariant if for every two families F and G and for every number $\lambda > 0$, the following two conditions are true:*

- *if F is better than G in the sense of this criterion (i.e., $F > G$), then $S_\lambda(F) > S_\lambda(G)$;*
- *if F is equivalent to G in the sense of this criterion (i.e., $F \sim G$), then $S_\lambda(F) \sim S_\lambda(G)$.*

Theorem 1. *If a family F is optimal in the sense of some optimality criterion that is final and scale-invariant, then every function d from F is equal to*

$$d(N) = \frac{A \cdot N^\beta}{1 + A \cdot N^\beta}. \qquad (1)$$

for some A and $\beta > 0$.

The proof is similar to the proofs presented in [2, 3].

Comment. According to the above formula, even if the number of occurrences N is very large, the degree of relevance is close to 1 but still not equal to 1. It may be reasonable to require that in such situations, we should get the degree of relevance equal to 1. With this is mind, we may want to have a function $d(N)$ which is only piece-wise smooth, with a smooth part continuously blending into identical 1 for large n; when applying the fractional-linear transformations, we shall take this "equal to 1" part into consideration. One can see from the proof that in this case, we end up with the following more general formula:

$$d(N) = \min\left(\frac{A \cdot N^\beta}{1 + B \cdot N^\beta}, 1\right),$$

for some $A \geq B$ and β.

Acknowledgments. This work was supported in part by NASA under cooperative agreement NCC5-209, by NSF grants No. DUE-9750858 and CDA-9522207, by United Space Alliance, grant No. NAS 9-20000 (PWO C0C67713A6), by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant number F49620-95-1-0518, and by the National Security Agency under Grant No. MDA904-98-1-0561.

References

[1] L. T. Koczy, T. D. Gedeon, and J. A. Koczy, *Proc. 8th IEEE Int'l Conf. on Fuzzy Systems (FUZZ-IEEE'99)*, Seoul, Korea, August 22–25, 1999, Vol. 1, pp. 158–163.

[2] H. T. Nguyen and V. Kreinovich, *Applications of continuous Mathematics to Computer Science*, Dordrecht: Kluwer, 1997.

[3] R. Osegueda, Y. Mendoza, O. Kosheleva, and V. Kreinovich, *Proc. 14th IEEE Int'l Symposium on Intelligent Control/Intelligent Systems and Semiotics ISIC/ISAS'99*, Cambridge, Massachusetts, September 15–17, 1999, pp. 208–212.