

Fuzzy Indices of Document Reliability

Célia da Costa Pereira¹ and Gabriella Pasi²

¹ Università degli Studi di Milano, DTI
Via Bramante 65, 26013 Crema (CR), Italy
pereira@dti.unimi.it

² Università degli Studi di Milano Bicocca, DISCO
Via Bicocca degli Arcimboldi 8, 20126 Milano (MI), Italy
pasi@disco.unimib.it

Abstract. This paper presents a first step toward the formalization of the concept of document reliability in the context of Information Retrieval (and Information Filtering). Our proposal is based on the hypothesis that the evaluation of the relevance of a document can also depend on the concept of reliability of a document. This concept has the following properties: (i) it is user-dependent, i.e., a document may be reliable for a user and not reliable for another user; (ii) it is source-dependent, i.e., the source which a document comes from may influence its reliability for a user; and (iii) it is also author-dependent, i.e., the information about who wrote the document may also influence the user when assessing the reliability of a document.

1 Introduction

The problem of information overload on the Web leads to a demand for effective systems able to locate and retrieve information relevant to user's individual interests. Usual systems for the content-based access to huge information repositories produce a ranked list of documents to be presented to the user. The rank is made possible by the estimate, by the system, of the so called "retrieval status value" (RSV). The usual evaluation criterion is relevance assessment, based on a formal representation and comparison of documents contents and the user's query content. In this case the RSV represents the system's estimate of the relevance of a document to the user's information needs. However, several additional properties of documents could be considered to assess their RSVs to users' needs [12]. Among these criteria, the *reliability* is a quite interesting one, which we try to approach and analyze in this paper. We present both a first analysis and some possible formalizations of the document reliability concept. Starting from a philosophical standpoint, we propose a fuzzy formalism for taking into account some different dimensions which concern the reliability of a document with respect to a particular user. More precisely, we attempt to give formal answers to the following questions:

- how the trust of the user in the source from which a document comes from, may determine document reliability?

- may the document reliability concept be corroborated by other sources?
- might other information like the author of the document and the date in which the document was been published contain useful information to enhance evaluation of document relevance?

To answer these questions, we have taken an inspiration from a philosophical approach [6]. We propose then some fuzzy indices which may be combined in distinct ways and which can be used in the process of document ranking.

The paper is organized as follows. In Section 2, we propose two fuzzy indices for evaluating the reliability of a document from the trust the user has for its source. In Section 3, we propose a third index for evaluating the reliability of a document. Thanks to this index, the case in which the information provided by the document is shared by other sources is also considered. In Section 4, we propose two other criteria as possible components in the process of evaluating document relevance. In section 5, we propose a simple way to combine the abovementioned indices into a unique index representing the overall reliability of a document. Finally, Section 6 concludes.

2 Evaluating the Reliability of a Document from the User Trust of Its Source

There are two different dimensions in the process of evaluating the user trust of a source. The first one, more intuitive and easily tractable, is when the user has had in the past a significant amount of relevant information from a given source. The second dimension is when the user either does not know at all or does not dispose of enough information to be allowed to evaluate the source. We will approach in a separate way either situation in the next sections.

2.1 When the Source is Known

In this case the user has a clear idea of the contents of the source, and she/he has been able to judge the information coming from that source as relevant. Therefore, when the source is *known*, the user can associate a trust degree with it; as a consequence the reliability degree of a document for the user i , noted $\mathcal{T}_i^k(d)$, may be computed in the basis of the degree to which the user trusts the source from which the document d comes from, i.e.,

$$\mathcal{T}_i^k(d) = T_i(s(d)), \quad (1)$$

where index k stands for *known* source, $s(d)$ represents the source of document d , and $T_i(s(d))$ represents the trust degree of the source for user i .

In the case in which the user may dispose of information describing the preferences of other users (such in the case of some collaborative filtering approaches) it would also be judicious to base the document reliability evaluation on the opinion of other users the user trusts. Thus doing, we may suppose that the user explicitly associates with a source a trust degree which depends also (i.e. it is

influenced) on the ones specified by the colleagues she/he trusts. The degree of trust a user i has in document d , originated from a known source, $\mathcal{T}_i^k(d)$, may in this case be given by:

$$\mathcal{T}_i^k(d) = \gamma * T_i(s(d)) + (1 - \gamma) * \mathbb{T}(s(d)) \quad (2)$$

where $\gamma \in]0, 1]$ is the degree of self trust of the user and $\mathbb{T}(s(d))$ represents the average degree on source $s(d)$ for all the users who user i trusts. It may be given by:

$$\mathbb{T}(s(d)) = \begin{cases} \frac{\sum_{j \in Users} t(i,j) * T_j(s(d))}{\sum_j t(i,j)} & \text{if } \exists j \text{ such that } t(i,j) \neq 0, \\ 0 & \text{Otherwise,} \end{cases}$$

where $t(i,j)$ represents the degree of trust user i has for user j .

2.2 When the Source is Unknown

Things are more complicated in the case in which the Web is the considered document repository, or in the case in which the information source is unknown. In fact, there is a huge number of information sources in the Web and, in most cases, when using for example a search engine, a user obtains information from quite different and "unknown" sources (sources she/he sees for the first time). By information sources we may here intend Web sites. Thus, it would be not possible to evaluate the user trust of a source "seen" for the first time; in this case "past track records" do not exist. When information is provided by Internet sources or other sources which do not have a well-established reputation, a possible approach for evaluating user trust is link analysis, i.e. to determine whether the information source is "strongly" endorsed by others by looking at how many Web sites link to that Web site. In the case of other kinds of unknown sources, we may act in an analogous way, depending on the type of considered documents. For example in the case of information source containing scientific documents the impact of a document could be evaluated by using citation counts. A possible model of this interpretation is inspired by the formula used by *Citeseer* [8] for calculating the impact of a scientific article in a given year. Let \bar{n} be the average number of citations for each document published in a given period, and n be the number of citations for the document d , published in that period. We define the impact of d , $\text{impact}(d)$, as

$$\text{impact}(d) = \log_2\left(\frac{n}{\bar{n}} + 1\right). \quad (3)$$

The impact index of d , $\mathcal{I}(d)$, is given by:

$$\mathcal{I}(d) = 1 - 2^{-\text{impact}(d)} \quad (4)$$

The degree of reliability of document d for user i is then given by:

$$\mathcal{T}_i^u(d) = \mathcal{I}(d), \quad (5)$$

where index u stands for *unknown* source.

The above proposition for computing the impact of a document does not take into account the “weight” of each citation. In fact, a citation which comes from a much cited document counts like the citation which comes from a little cited document and this is not reasonable. To take the “weight” of citations into account, we propose a function inspired by the notion of *PageRank* proposed by [2] for ranking pages on Google, which makes the calculation of the weighted value of a citation possible.

Let d_1, \dots, d_m be m documents which cite a document d , and $o(d_i)$ the number of citations pointing out from d_i . The *weighted value of citations* of d , $n(d)$, may be computed by using the following iterative procedure:

$$n(d) = \frac{n(d_1)}{o(d_1)} + \dots + \frac{n(d_m)}{o(d_m)} \quad (6)$$

where $n(d_i)$ is the weighted value of citations to document d_i . The intuitive justification of this formula is that a document has a higher citation value if many other documents point to it. This value may increase if there are high-scoring documents pointing to it.

3 When Information Is Corroborated Can It Help in Evaluating the Reliability of a Document?

In the previous section, we looked at ways to determine the reliability of a document by considering a single source of information. In this section, we attempt to evaluate the reliability of a document by also considering the case in which the information provided by that document is shared by other sources. In fact,

in addition to the character of the “witness”, we may pay attention to the “number of witnesses”. This is because it is much more likely that one individual will “deceive or be deceived” than that several individuals will “deceive or be deceived” in exactly the same way [6]

(a remarkable counterexample is when the majority of the electors of a country is deceived by the false promises of a politician).

For this reason, several philosophers have noted that the agreement of a number of experts on a topic can be an important indicator of accuracy. This suggests that another technique for verifying the accuracy of a piece of information is to see if other information sources corroborate the original source of the information [6].

Of course,

information sources do not always agree with each other. In fact, it is fairly easy to find conflicting information. If sources do conflict, then people simply have to determine which source is more reliable [6]

(or use some of the other techniques for verifying the accuracy of the information).

Notably, however, agreement between information sources is not always an indication that their information is accurate. It depends on how these different sources got their information. In particular, if they all got their information from the same place, then ten sources saying the same thing is no better evidence than one source saying it. This issue turns out to be especially important on the Internet since it is so easy for the very same information to be copied by several different Web sites. However, the fact that all of these sites corroborate each other still does nothing to help us verify that the information is accurate. Agreement between sources should not increase our degree of confidence in the accuracy of a piece of information unless those sources are independent. Obviously, here what is required is only a *conditional independence* and not full independence. In fact, if two information sources are reliable, their reports will be correlated with each other simply because their reports will both be correlated with the truth. Thus, it is very difficult to calculate the reliability of a piece of information based on the fact that it is or not corroborated by other sources without making some trade-off [6].

Here, we propose to evaluate the accuracy of a piece of information by checking if such information is shared by other sources.

Let d be a new document and d' be a document obtained from source s' which is such that:

- d' is the document more similar to document d , i.e. $\neg\exists d''$ with $s(d'') = s' \neq s(d)$ such that $\text{sim}(d, d'') > \text{sim}(d, d')$ ¹;
- $\text{sim}(d, d') \geq \alpha$, where α is the similarity degree after which two documents are considered similar.

The degree of trust a user i has for document d , $\mathcal{T}_i^c(d)$, may be defined as follows:

$$\mathcal{T}_i^c(d) = \frac{\sum_{s' \in \text{Sources}} \mathcal{T}_i(s') * \text{sim}(d, d')}{\sum_{s'} \mathcal{T}_i(s')}, \quad (7)$$

where $\text{sim}(d, d')$ is set to 0 if $\text{sim}(d, d') < \alpha$, i.e., we consider only documents similar to d . This allows us to take also into account the number of sources sharing documents similar to d . Another possibility is

$$\mathcal{T}_i^c(d) = \max_{s' \in \text{Sources}} \min(\mathcal{T}_i(s'), \text{sim}(d, d')) \quad (8)$$

In both cases, if there is no source in which the user trusts, or if there is no source containing at least a document more or less similar to document d , the user concludes that the information in document d is not corroborated by any source she/he trusts and, consequently, she/he does not trust d .

¹ Sim is a function which returns the similarity degree between two documents.

4 Other Information Which Could Enhance Evaluation of Document Relevance

In this section, we propose two other criteria, namely the author of the document and the publication date of the document, as possible components in the process of evaluation of document relevance. In fact, in the case in which the source contains, for example, scientific documents, information about who wrote the document may be useful for determining the relevance of the document. In a similar way, if the piece of information has a temporary “truth value”, e.g., it is merely relevant during a certain period, the date in which the document was written may strongly contribute to evaluating its relevance.

4.1 Information About the Author of the Document

As we have said previously, information about the author of a document may also be useful to determine the reliability of the document for a user. This information may include:

- the author’s name, N_a ($\in \{0, 1\}$). More precisely, $N_a = 1$ if the document contains the name of the author and 0 otherwise;
- the author’s education or experience, E_a ($\in [0, 1]$) which represents how the author experience is relevant to the interest group of the document. The calculation of this degree may be made based on a comparison between the interest theme of the document and the interest theme of the author’s qualification, degree, or scholarship on one hand and, on the other hand, it may be based on an evaluation made by the user of past documents written by the same author;
- contact information, I_a ($\in \{0, 1\}$). More precisely, $I_a = 1$ if the document contains a contact information of the author and 0 otherwise;
- the name of the organization for which she/he works, O_a ($\in \{0, 1\}$). More precisely, $O_a = 1$ if the document contains the name of the author’s organization and 0 otherwise.

The following importance order may be established according to common sense: $N_a = I_a > E_a > O_a$. This order means that the most important piece of information is the name of the author and her/his contact information, followed by the author experience and the least important piece of information is the organization for which she/he works. Therefore, the trust degree of document d for user i , considering information about the author, may be calculated as follows:

$$\mathcal{T}_i^a(d) = \lambda_1 \cdot (N_a + I_a)/2 + \lambda_2 \cdot E_a + \lambda_3 \cdot O_a \quad (9)$$

with $\lambda \in [0, 1]$ and $\sum_{i=1}^3 \lambda_i = 1$ and $\lambda_1 > \lambda_2 > \lambda_3$.

4.2 Date of the Document

The presence or absence of the date ($D \in \{0, 1\}$) on which a document was written may also help in determining its reliability. For some themes, the information contained in a document with an old date may become irrelevant/obsolete/not

true. For example, in the sport theme, information that “Juventus plays in premier league, 2005”, is not reliable in 2006. Instead, information that “Dante Alighieri was born in Florence in 1265”, is always reliable. Thus, an old date on information known to be changeable is a sign of irrelevance.

We propose to take this fact into account by combining all user’s interests into two groups, namely, those in which the date influences their relevance and those in which it does not. The reliability degree of document d for author i , considering the presence or absence of the date in which the document was written, may be calculated as follows:

$$\mathcal{T}_i^d(d) = (1 - \beta) + \beta * D. \quad (10)$$

$\beta \in [0, 1]$ is high if the document belongs to the first group, and low otherwise.

5 Combining Reliability Degrees

Each of the user trust degrees proposed in the previous sections corresponds to different trust degrees the user may give to a document. We propose to combine all of these degrees to obtain the overall *degree of reliability* of a document d for a user i , $\mathcal{R}_i(d)$, as follows:

$$\mathcal{R}_i(d) = \delta_1 * \mathcal{T}_i^j(d) + \delta_2 * \mathcal{T}_i^c(d) + \delta_3 * \mathcal{T}_i^a(d) + \delta_4 * \mathcal{T}_i^d(d), \quad (11)$$

$j = k$ if the source is known otherwise $j = u$; $\delta_i \in [0, 1]$ and $\sum_i^4 \delta_i = 1$. Of course, the above overall *degree of reliability* will then have to be combined with some conventional document-similarity measure and other criteria to obtain the RSV of a document.

6 Conclusion

To answer a number of questions stated in the Introduction, an extensive critical survey of the literature about relevance and trustiness, in particular in the philosophical domain, has been carried out. The material gathered has been elaborated on and formalized, resulting in the proposal of several fuzzy measurements of user trust which were combined to obtain the overall reliability of a document for a user. Since the relevance of a document for a user depends also on the reliability of that document for that user, we believe that this proposal may be useful in automated methods to locate and retrieve information with respect to individual user interests.

References

1. Alexander, J.E., Tate, M.A.: Web Wisdom; How to Evaluate and Create Information Quality on the Webb. Lawrence Erlbaum Associates, Inc, Mahwah, NJ, USA (1999)
2. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1–7), 107–117 (1998)

3. Bruce, B.: Credibility of the web: Why we need dialectical reading. *Journal of Philosophy of Education* 34, 97–109 (2000)
4. Burbukes, N.C.: Paradoxes of the web: The ethical dimensions of credibility. *Library Trends* 49, 441–453 (2001)
5. Dubois, D., Prade, H.: *Possibility Theory – An approach to Computerized processing of Uncertainty*. Plenum Press, New York (1988)
6. Fallis, D.: On verifying the accuracy of information: Philosophical perspectives. *Library Trends* 52(3), 463–487 (2004)
7. Ketelaar, E.: Can we trust information? *International Information and Library Review* 29, 333–338 (2004)
8. Ley, M.: Estimated impact of publication venues in computer science (2003)
9. Lynch, C.A.: When documents deceive: Trust and provenance as new factors for information retrieval in a tangled web. *Journal of the American Society for Information Science and Technology* 52(1), 12–17 (2001)
10. Matthew, R., Agrawal, R., Domingos, P.: Trust management for the semantic web (2003)
11. Tomlin, J.A.: A new paradigm for ranking pages on the world wide web. In: *WWW '03. Proceedings of the 12th international conference on World Wide Web*, New York, NY, USA, pp. 350–355. ACM Press, New York, NY, USA (2003)
12. (Calvin) Xu, Y., Chen, Z.: Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol.* 57(7), 961–973 (2006)
13. Zadeh, L.A.: Fuzzy sets. *Information and Control* 8, 338–353 (1965)