Implication in Information Retrieval Systems

Laurent Ughetto
IRISA - Rennes 2 University
Rennes, France
laurent.ughetto@irisa.fr

Gabriella Pasi
Università degli Studi di Milano Bicocca
Milano, Italia
pasi@disco.unimib.it

Vincent Claveau
IRISA - CNRS
Rennes, France
vincent.claveau@irisa.fr

Olivier Pivert
IRISA - ENSSAT
Lannion, France
olivier.pivert@enssat.fr

Patrick Bosc
IRISA - ENSSAT
Lannion, France
patrick.bosc@enssat.fr

ABSTRACT

Some IR models make use of an implication to match a document d and a query q, computing either "q implies d" (e.g. in fuzzy inclusion models) or, the other way, "d implies q" (e.g. in logical IR models). This paper analyzes, from a theoretical point of view, the IR models using both approaches. Even if the above notations seem to be opposite, it is shown that they sometimes come from different formulations of the same paradigm, which led to mistakes in the literature. Then the paper comes back to fuzzy models based on "q implies d" (q included in d) and shows their efficiency, and compares them to models based on "d implies q" (d included in q). The latter is attractive from a theoretical point of view, but turns out to be less efficient in practice, and is rarely adopted in the literature. At last, attempts to use "d implies q" in a fuzzy model are discussed, and we try to explain their inefficiency.

Keywords

IR models, fuzzy logic, fuzzy implication

1. INTRODUCTION

Since the work by van Rijsbergen [14] in 1986, several Information Retrieval (IR) models have used an implication to determine if a document d is relevant to a query q. Sometimes, this implication is used to model the inclusion of single query terms in a document (**if** the term is in the query, **then** it should be in the document). This inclusion is denoted by $q \rightarrow d$ [12, 3]. In logical IR models, the relevance of a document to a given query is modeled by the notion of logical consequence: the query should be a logical consequence of a relevant document. This is denoted by $d \rightarrow q$ [7]. Although the above approaches are the main ones using an implication, others have appeared in the literature. Depending on the considered implication, logic and kind of relevance, one can find either the notation $d \rightarrow q$ or $q \rightarrow d$ in the literature. Interestingly, the two notations, which seem to be diametrically opposed, have been implemented in systems whose functioning does not always seem so different.

From these observations, the main goal of this paper is to list and study, from a theoretical point of view, the different uses of implication in the IR literature. It is interesting to understand the basis of the different approaches, the different ways an implication is used, and to point out their shared properties and differences. Thus, the different formalisms using an implication are briefly exposed and explained.

Usually, the two aforementioned notations are used in different approaches. For instance, when the implication is used to represent set inclusion, the inclusion degree of a query in a document, or of a document in a query are two different, complementary approaches to determine the relevance of a document. However, more surprisingly, these two notations $d \rightarrow q$ and $q \rightarrow d$ sometimes represent equivalent approaches, just modeled differently. As a bad consequence, the same notation sometimes represents different approaches, which leads to confusion, and sometimes to mistakes, as in [11].

At last, this paper comes back to the fuzzy extension of the Boolean model developed in [12, 3], where the implication is used to compute the inclusion degree of a query in a document, whose efficiency has been shown in [1]. Then, the last part tries to show why the dual approach, which consists in computing an inclusion degree of a document in a query is more tricky to exploit, and has not produced an efficient IR model despite its theoretical attractiveness.

2. IR LOGICAL APPROACHES

2.1 IR logical models in the literature

IR logical models have been studied by many authors during the 90s. Keith van Rijsbergen was among the first to propose a logical interpretation of information retrieval, using the concept of implication: a query is implied by a document, $d \rightarrow q$, where $\rightarrow$ is an implication operator from the considered logic [14]. Logical approaches are founded on the representation of both the documents and queries by logical formulae from the considered logic. Thus, the implication should not be the material implication from classical logic. This is an important point in the proposition of van Rijsbergen: the choice of the *good* operator depends on the logic used in the formalization process. He pointed out that a non-classical logic is probably more accurate to model the IR process. This is why, in his proposition, the implication of the query by a document is associated with a probability degree:

RIAO '2010 Paris, France

$P(d \rightarrow q)$. From this first work, several authors have studied the role logic may play in IR models. An overview, and a fine analysis of the different approaches in the literature have been proposed by Sebastiani [13] and Lalmas [7].

Logical models proposed in the literature differ on two main aspects: the logical state of the formula $d \rightarrow q$ used to determine if the document is relevant, and the logic $\mathcal{L}$ chosen.

As to the first aspect, the mechanism used to determine the relevance of a document d to a query q depends on the interpretation of the formula $d \rightarrow q$. As recalled in [13], the possible interpretations are:

- $d \rightarrow q$ is true in some particular interpretation of the chosen logic [10, 5],
- d is a logical consequence of q in the logic $\mathcal{L}$,
- the formula $d \rightarrow q$ is valid in $\mathcal{L}$ [8],
- q can be derived from d in $\mathcal{L}$ [15, 4],
- $d \rightarrow q$ is a theorem of $\mathcal{L}$.

Even if logical consequence, validity, derivability and theoremhood are equivalent in sound and complete logics, these notions may encompass rather different meanings, depending on the considered logic $\mathcal{L}$. Then, any IR model should take this semantics into account, and choose the way to compute the relevance of a document d accordingly. However, the notion of validity is the most widely used.

A second difference between the proposed approaches is the kind of logic chosen to model the IR process. For instance, there are propositions founded on modal logic, logical imaging, terminological logic [10, 6, 8]. Nie also proposed a metamodel based on modal logic, in order to redefine existent IR models using logic [9]. He considered the two approaches denoted $d \rightarrow q$ and $q \rightarrow d$. The former is linked to the concept of exhaustivity, and the latter to the concept of specificity. Numerous kinds of logic have been used (or at least proposed in theoretical studies), taking advantage of their various properties. The reader can refer to [7] for an overview of these logics, and their advantages in an IR model. Despite these numerous studies, IR logical models are seldom used in practice.

2.2 Principles of IR logical models

This section briefly explains the principles of IR logical models, using propositional logic for the sake of simplicity.

In this approach, documents and query are represented by logical formulae (hence the name), and most of the time a conjunction of the index terms they contain. For instance, a document d_i defined by the set of terms $\{t_1, t_2, t_3, t_5\}$ is represented by the formula $d_i = t_1 \wedge t_2 \wedge t_3 \wedge t_5$. Although it can be represented by a more general formula, a query q is often a conjunction of terms as in "bag of words"-like models.

In order to determine if d_i is relevant to q, a logical IR model checks the status of the formula $d_i \rightarrow q$. As stated above, it can be done in four ways, which are equivalent in propositional logic. The first ones come from model theory:

- $\vDash d_i \rightarrow q$: formula $d_i \rightarrow q$ is valid (i.e., true whatever the truth of terms t_j),
- $d_i \vDash q$: formula q is a logical consequence of d_i (i.e., valuations satisfying d_i also satisfy q).

The two others come from proof theory:

- $\vdash d_i \rightarrow q$: formula $d_i \rightarrow q$ is a theorem,,
- $d_i \vdash q$, formula q may be derived from formula d_i (using a proof method).

With other logics, more accurate in IR, these methods may not be equivalent (or even feasible). See [7] for examples.

3. SET-ORIENTED APPROACHES

3.1 Inclusion and Boolean models

Considering that documents and queries are sets of terms, inclusion can be seen as a simple IR method : a document is relevant if and only if it contains all the query terms. Let d_i, $i \in \{1, \ldots, n\}$ be the n documents from collection C, q the query, and $t \in T$ the index terms, then the relevance of document d_i is given by the inclusion formula:

$$\forall t \in T,\ (t \in q) \Rightarrow (t \in d_i)\ . \tag{1}$$

It corresponds to the division operation from relational algebra, as it answers the relational query: "what are the documents containing all the query terms?" (see [3] for details).

The first IR model, the well-known Boolean model, is founded on this inclusion model from set theory, and on Boolean Logic (in order to allow more general queries). In this model, a document is a set of words, but a query is a logical formula made of terms linked by AND, OR, NOT operators, which can be written in *disjunctive normal form.* A document is relevant iff, for at least one of the conjunctive clauses of the query, all the non-negated terms are included in the document and all the negated terms are not included in it.

This method has well-known drawbacks, as for instance: i) document and query terms cannot be weighted, and then their varying importance cannot be taken into account, ii) there is no inclusion *degree* of a query term in the document, iii) a document is judged irrelevant if just one required term is absent from the document (or if just one negated term is present), iv) as a consequence relevant documents cannot be ordered. To overcome these limitations, extensions have been proposed, using term weights and computing relevance degrees.

3.2 Gradual implication-based IRS

In [12], a new way to use logic in IR was proposed. In contrast with logical models, the implication $\rightarrow$ is viewed as the material implication. Based on the analysis of the Boolean model, this approach does not connect the entire document d_i to the entire query q, as logical models do, but it connects one term t from the query with the same term t from the document: $q(t) \rightarrow d_i(t)$. The correspondence between a query and a document, at the terms level, is assessed by the truth degree of the implication, with the interpretation given by the query and document (i.e., $q(t)$ is true if t belongs to query q, and $d_i(t)$ is true if t belongs to document d_i). The way the implication is written comes naturally from the modeling of the relevance concept in the Boolean model, where a document is relevant to a query if its representation contains the query terms.

Independently, this approach was proposed again, two years ago [3]. The authors had worked for several years on the division of fuzzy relations in a fuzzy databases framework (e.g., [2]). Then, noticing that the Boolean IR model is linked to the division of relations in databases, they envisioned that the division of fuzzy relations could correspond to a good IR model.

However, this model was just a theoretical proposition, and was not experimentally validated. It has been implemented, tested and validated last year [1].

While the Boolean model computes the inclusion of the query in the document (in case of conjunctive queries[1]), its fuzzy extension computes an *inclusion degree* which makes it possible to order the relevant documents. Queries and documents are represented by fuzzy sets on the universe of index terms T. For instance, a query q is represented by the set $\{\alpha_j/t_j,\ j = 1, \ldots, m\}$, where α_j/t_j means that the term t_j belongs to the query to the degree α_j. This degree, or weight, α_j is often denoted by $\mu_q(t_j)$, the membership degree of term t_j to the query q. Similarly, a document d_i is a fuzzy set on T, and $\mu_{d_i}(t_j)$ if the membership degree of term t_j to this document. These membership degrees correspond to the weights of the queries and documents terms in usual IR models, but they belong to the unit interval $[0, 1]$ (as membership grades do), while term weights are more general real values.

The formula that describes the computation of the inclusion degree is a straightforward extension of expression (1):

$$\mathrm{Inc}_q(d_i) = \top_{t \in q} \left(\mu_q(t) \to \mu_{d_i}(t)\right) \ , \qquad (2)$$

where $\top$ is a t-norm (a fuzzy conjunction) and $\to$ is a fuzzy implication. There is a close link between this formula, and score formulae in classical IRSs like Okapi, as:

- the fuzzy implication $\to$ corresponds to the matching function between the weight of a term in a query, and its weight in the document, used to compute an individual term score;
- the t-norm $\top$ corresponds to the aggregation function computing a document score from the terms scores.

The experimental validation reported in [1] has shown the importance of well choosing the fuzzy matching and aggregation operators, among the wide range of available fuzzy operators. First, the semantics of the term weights in the query depends on the kind of fuzzy implication (importance with S-implication, and threshold with R-implications). Second, the implication and t-norm must have good properties in order to obtain an efficient system, competing with Okapi.

4. QUERY INCLUDED IN DOCUMENT

In most IR models based on an implication, the relevance degree is an inclusion degree of the query in the document.

In the previous sections, two kinds of IR approaches, both founded on logic, have been presented (the set-based models as the Boolean model, and the logical models). At first glance, these approaches seem opposed, as they lead to opposite formalizations: $q \to d_i$ or $d_i \to q$. However, it can be shown that it is just a matter of notation, and that the logical IR model based on propositional logic corresponds to the Boolean IR model. The difference is due to the formalization process, in which d_i and q do not represent the same thing in the two approaches. This equivalence can be formally shown. However, for the sake of brevity, it is just shown intuitively for conjunctive queries.

In the Boolean model, a document is relevant if it contains all of the query terms. In the logical model a document is relevant if for each valuation satisfying d_i (i.e., when all the document terms are true) formula q is also true. When q is a conjunction of the terms it contains, it is true only if all its terms are true, and for that they have to be in the document. This means, as in the Boolean model, that the query terms must be in the document.

Formulas differ, but the condition is the same. If $d_i \to q$ is a good notation for logical models, where d_i and q represent the entire document and query, $q(t) \to d_i(t)$ (and maybe $\forall t,\ q(t) \to d_i(t)$) would be better for set-based models, where implication is at the terms level. It will be used from now on.

If the Boolean model is inefficient, its fuzzy extension can compete with Okapi, when the operators and weights are well chosen [1].

5. DOCUMENT INCLUDED IN QUERY

Only few studies have aimed at extending the paradigm $d_i \to q$, in the sense of $d_i \subseteq q$. This section first reviews the work by Oussalah *et al.* [11], then our implementation of the approach. It also tries to show why they yield bad results.

5.1 A debatable approach

In the paper entitled "Personalized information retrieval system in the framework of fuzzy logic", Oussalah *et al.* claim to implement an IR search model based on the paradigm $d_i \to q$, in the sense of $d_i \subseteq q$.

This work is interesting as it is one of the very few studies trying to do so. It also provides interesting insights on the use of fuzzy logic in IR. Unfortunately, it seems that the authors have adopted this approach by bundling the notations $d_i \to q$ and $q(t) \to d_i(t)$. The implication-based formula at the heart of their system has undergone an *ad hoc* modification leading to a model different from that announced, and was compared to a Boolean system only.

Starting from the Boolean model, their work tries to extend it using fuzzy logic. As in the Boolean model, the document is considered as a set of terms, and a query is represented as a set of index terms connected by logical operators like AND, OR, NOT. However, it is explained that: "[...] a document d answers a query q, if the implication $D \Rightarrow Q$ holds, where D and Q stand for some logical formula of document d and query q [...]" which clearly references logical models, in which queries and documents are logical formulae, and Q has to be satisfied when D is.

Then, the authors implement the paradigm $d_i(t) \to q(t)$, which computes an inclusion degree of a document in the query. The degrees $d_i(t) \to q(t)$ are computed for each term t, but are aggregated with a triangular conorm (a fuzzy OR), specifically the *max*, while a real inclusion would rather use a triangular *norm* (a fuzzy AND), as in formula (2). The authors note, rightly so, that when a term t is absent from the document and the query, the implication equals 1. But since 1 is the absorbing element of any t-conorm, the whole document is thus considered fully relevant! Thus, they proposed an *ad hoc* modification which reduces to replacing $d_i(t) \to q(t)$ with $\mathit{min}(d_i(t), q(t))$, i.e. to replacing the implication with a conjunction. And, for the sake of efficiency, the aggregation operator *max* is replaced by the fuzzy algebraic sum, which allows for a better ordering of relevant documents.

At last, their model is far from the announced paradigm.

5.2 Inconclusive attempts

This section reports some experiments related to our attempt to implement the approach $d_i(t) \to q(t)$ by inverting

[1] Without a loss of generality, only bag-of-words-like conjunctive queries are considered from now on, in order to simplify notations and explanations.

$q(t)$ and $d_i(t)$ in formula (2). Such an approach computes the inclusion degree of the document in the query. It seems obvious that the bigger the part of the document in the query, the more relevant the document. However, a straight implementation, using the same conditions described in [1], and just inverting the document and query in the formula, yields very bad results: from -80% to -100% when compared with Okapi (with the standard parameters).

Trying to understand why such bad results are obtained compared to the dual approach is interesting. First, let us examine the case $q(t) \rightarrow d_i(t)$. According to formula (2), the scores of individual terms are computed and aggregated with a fuzzy AND (a "good" one has to take into account every term in the computation of the final score). Moreover, a document term absent from the query gets the maximal score 1, which is the neutral element for the aggregation operator. Thus, such terms do not contribute to the final score, which can be computed from the terms in the query only. Among these terms, those having a high weight in the document receive a high individual score, close to 1, the neutral element for the aggregation (to simplify, query term-weights are assumed to be 1 here). Conversely, terms from the query absent from the document or present with a small weight receive a small score close to 0. Due to the use of a t-norm, low-weight terms have a stronger impact on the aggregation than those with a high weight.

For example, let us consider a query of three terms, two documents, and their respective individual term scores $d_1 = \{1,\ 1,\ 0.1\}$, $d_2 = \{1,\ 0.4,\ 0.3\}$. With the t-norm product, the final score of d_1 is 0.1 whereas that of d_2 is 0.12.

This behavior is opposed to the aggregation as performed by efficient systems like BM-25 vector-space models in which d_1 would yield a better score than d_2. Yet, if fuzzy-logic-based systems perform well, it is due to the fact that documents are compared on the same set of query terms. As the number of query terms is constant, the number of these terms present in the document is inversely proportional to the number of the absent ones. Then, measuring the part of the query outside the document, or the part of the query inside the document lead to the same documents ranking.

Systems based on $d_i(t) \rightarrow q(t)$ work similarly. Document terms absent or only slightly present in the query are those having the strongest impact on the final score. But in this case, the terms intervening in the score computation are those from the document, not from the query. These terms are numerous, and their number vary according to the document. While we would like to measure the (fuzzy) quantity of document terms which are also in the query, we actually measure the quantity of terms which are not in the query, which also depends on the size of the document. A first idea to solve this issue is to normalize the size of the document. The (simple) normalization techniques (number of terms, sum of weights...) which have been tested improved the results but not enough. Yet, several other techniques are still to be explored, but the structural issues described above make this approach not as straightforward as the $q(t) \rightarrow d_i(t)$ approach.

6. CONCLUSION

This paper presented an overview on the use of implication in IR models. It has been shown that notations which seem to be opposed are sometimes only different formulations of the same paradigm. Indeed, implications $d \rightarrow q$ from logical models, and $q(t) \rightarrow d(t)$ from (fuzzy) set-based models, both mean that query terms have to be included in the set of terms associated with a document, for this document to be relevant. This approach has been implemented in efficient IR systems.

The opposite inclusion, $d(t) \rightarrow q(t)$, although mentioned in the literature, has rarely been used in practice (and sometimes in a confused way). Even though it has not been efficiently implemented, its underlying principle is still appealing. It could be complementary to the $q(t) \rightarrow d(t)$ approach, either by combining the results of both models, or using a bipolar approach. Indeed, the inclusion degree of q in d can be interpreted as a possibility degree for d to be relevant, and the inclusion degree of d in q as a necessity degree. This is why the $d(t) \rightarrow q(t)$ paradigm still deserves our attention.

Last, it has been shown that the Boolean model is equivalent to the logical model when using propositional logic. However, the fuzzy extensions of these two models may be not equivalent. Fuzzy-logic-based extensions of the logical models could lead to another interesting, and efficient IR model.

7. REFERENCES

[1] P. Bosc, V. Claveau, O. Pivert, and L. Ughetto. Graded-inclusion-based information retrieval systems. In *Proc. of ECIR'09*, pages 321–336, 2009.

[2] P. Bosc, D. Dubois, O. Pivert, and H. Prade. Flexible queries in relational databases – the example of the division operator. *Theor. Comp. Sc.*, 171:281–302, 1997.

[3] P. Bosc and O. Pivert. On the use of tolerant graded inclusions in information retrieval. In *Actes de CORIA'08*, pages 321–336, 2008.

[4] P. Bruza and L. van der Claag. Efficient context-sensitive plausible inference for information disclosure. In *Proc. of SIGIR'93*, pages 12–21, 1993.

[5] Y. Chiaramella and J.-P. Chevallet. About retrieval models and logic. *The Computer J.*, 35:233–242, 1992.

[6] F. Crestani and C. van Rijsbergen. Information retrieval by logical imaging. *J. of Doc.*, 51:293–331, 1995.

[7] M. Lalmas. Logical models in information retrieval: Introduction and overview. *IPM*, 34(1):19–33, 1998.

[8] C. Meghini, F. Sebastiani, U. Straccia, and C. Thanos. A model of information retrieval based on a terminological logic. In *Proc. of SIGIR'93*, pages 298–307, 1993.

[9] J.-Y. Nie. An outline of a general model for information retrieval systems. In *Proc. of SIGIR'88*, pages 495–506, 1988.

[10] J.-Y. Nie. An information retrieval model based on modal logic. *IPM*, 25(5):477–491, 1989.

[11] M. Oussalah, S. Khan, and S. Nefti. Personalized information retrieval system in the framework of fuzzy logic. *Expert Syst. with Appl.*, 35:423–433, 2008.

[12] G. Pasi. A logical formulation of the Boolean model and of weighted Boolean models. In *LUMIS workshop at ECSQARU'99*, 1999.

[13] F. Sebastiani. On the role of logic in information retrieval. *IPM*, 34(1):1–18, 1998.

[14] C. J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29(6):481–485, 1986.

[15] C. J. van Rijsbergen. Towards an information logic. In *Proc. of SIGIR'89*, pages 77–86, 1989.