Pergamon

Mathl. Comput. Modelling Vol. 27, No. 2, pp. 61–79, 1998
Copyright©1998 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0895-7177/98 $19.00 + 0.00

PII: S0895-7177(97)00260-4

Properties of Topologies of Information Retrieval Systems

L. Egghe
LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium
legghe@luc.ac.be

(Received and accepted April 1997)

Abstract—This paper studies topological properties of different topologies that are possible on the space of documents as they are induced by queries in a query space together with a similarity function between queries and documents. The main topologies studied here are the retrieval topology (introduced by Everett and Cater) and the similarity topology (introduced by Egghe and Rousseau).

The studied properties are the separation properties T_0, T_1, and T_2 (Hausdorff), proximity and connectedness. Full characterizations are given for the diverse topologies to be T_0, T_1, or T_2. It is shown that the retrieval topology is not necessarily a proximity space, while the similarity topology and the pseudo-metric topology always are proximity spaces. A characterization of connectedness in terms of the Boolean NOT-operator is given, hereby showing the intimate relationship between IR and topology.

Keywords—Topology, IR-system, Similarity, Separation, Connectedness, Proximity.

1. INTRODUCTION

In [1] (see, also, [2] for a correction of this paper), the retrieval topology is defined on a set of documents as follows. Let DS be the set (space) of all documents and QS a set of queries. In Everett and Cater, QS consists of all possible queries; in our vision (see [3]) QS consists of only elementary requests, i.e., not consisting of any Boolean (or other) combinations. Right now this difference is not important: QS is just a set of queries. Let $\mathrm{sim}(.,Q)$ be a function on DS into $\mathbb{R}$, measuring the similarity $\mathrm{sim}(D,Q)$ between a document D and a query Q. Most commonly, its values are in $\mathbb{R}^+$ or even [0,1], but this restriction is not necessary. The retrieval topology τ on DS is defined to be the topology generated by the sets (as a subbasis)

$$R(Q,r) = \{D \in DS \,\|\, \mathrm{sim}(D,Q) > r\}, \tag{1}$$

for $Q \in QS$ and $r \in \mathbb{R}$. These sets are called the retrievals of the system $(DS, QS, \mathrm{sim}, \tau)$ or (shortly) of τ. It is linked to the idea of the retrieval of documents via a threshold requirement on the value of the similarity measure. The set

$$\mathrm{ret}_\tau(Q) = \{R(Q,r) \,\|\, r \in \mathbb{R}\} \tag{2}$$

is called the set of retrievals of $Q \in QS$, or shortly the retrievals of Q (w.r.t. τ).

Also in [1], the pseudo-metric topology τ' on DS is defined as the topology generated by the subbasis

$$V(D,\varepsilon) = \{E \in DS \,\|\, |\mathrm{sim}(D,Q) - \mathrm{sim}(E,Q)| < \varepsilon,\ \forall Q \in QS\}, \tag{3}$$

Typeset by $\mathcal{A}_\mathcal{M}\mathcal{S}$-TEX

for $\varepsilon > 0$. There are several pseudo-metrics that generate τ'. One of them can be given by

$$d(D,E) = \sup_{Q \in QS} |\operatorname{sim}(D,Q) - \operatorname{sim}(E,Q)|, \tag{4}$$

if the sup is finite. If it is infinite, we can use the pseudo-metric

$$d'(D,E) = \sup_{Q \in QS} \frac{|\operatorname{sim}(D,Q) - \operatorname{sim}(E,Q)|}{1 + |\operatorname{sim}(D,Q) - \operatorname{sim}(E,Q)|} \tag{5}$$

(see [3]).

Finally, in [3], the similarity topology τ'' on DS is defined as the topology on DS which makes the similarity functions

$$\begin{aligned} \operatorname{sim}(.,Q) &: DS \to \mathbb{R}, \\ D &\mapsto \operatorname{sim}(D,Q), \end{aligned}$$

continuous. This topology τ'' is generated by the sets (as subbasis)

$$U(Q,r_1,r_2) = \{D \in DS \parallel r_1 < \operatorname{sim}(D,Q) < r_2\}, \tag{6}$$

$Q \in QS$, $r_1, r_2 \in \mathbb{R}$, $r_1 < r_2$. These sets are called the retrievals of the system $(DS, QS, \operatorname{sim}, \tau'')$ or (shortly) of τ''. It is linked to the idea of retrieval of documents via a "close match" of similarity values. The set

$$\operatorname{ret}_{\tau''}(Q) = \{U(Q,r_1,r_2) \parallel r_1 < r_2\} \tag{7}$$

is called the set of retrievals of $Q \in QS$ or shortly the retrievals of Q (w.r.t. τ'').

It is proven in [3] that $\tau \subset \tau'' \subset \tau'$, and (by example) that strict inclusions can hold (i.e., the three topologies can be different). It is also proved there that $\tau' = \tau''$ iff the set

$$\{\operatorname{sim}(.,Q) \parallel Q \in QS\}$$

is equicontinuous (e.g., when QS is finite) ([3, Theorem I.11]).

For more information on the topological terminology and properties, we refer the reader to Appendix A in [3] or to the vast literature on topological spaces, e.g., [4–8].

The following definition is taken from [1]: we say that an IR model $(DS, QS, \operatorname{sim})$ separates the points of DS if from $\operatorname{sim}(D,Q) = \operatorname{sim}(E,Q)$, $\forall\, Q \in QS$ it follows that $D = E$. In this case, τ' is a metric topology.

It is shown in [3] that τ as well as τ'' can be considered as the sets of all possible Boolean retrievals (using AND and OR) based on queries $Q \in QS$. Let us explain this for τ. Let $(Q_{ij})_{i=1,j=1}^{n \;\; m}$ be an array of queries in QS. The Boolean query denoted by

$$\underset{j=1}{\overset{m}{\mathrm{OR}}} \left(\underset{i=1}{\overset{n}{\mathrm{AND}}}\, Q_{ij} \right) \tag{8}$$

is defined through its set of retrievals

$$\operatorname{ret}\left(\underset{j=1}{\overset{m}{\mathrm{OR}}} \left(\underset{i=1}{\overset{n}{\mathrm{AND}}}\, Q_{ij} \right) \right) = \left\{ \bigcup_{j=1}^{m} \left(\bigcap_{i=1}^{n} R(Q_{ij}, r_{ij}) \right) \parallel r_{ij} \in \mathbb{R} \right\}. \tag{9}$$

It can be shown [3] that (8) represents a general Boolean query and that the sets in (9) for all $Q_{ij} \in QS$ $(m, n \in \mathbb{N})$ form the topology τ on DS when DS is finite. The same can be said for τ'' now using the sets $U(Q, r_1, r_2)$.

This shows the intimate relationship between (Boolean) IR and topology on DS.

In the sequel, we will show that the Boolean NOT-operator also plays an important role in the topological properties of DS.

This is what this paper is all about: showing some topological properties of the spaces (DS, τ), (DS, τ''), and (DS, τ'), if possible by using IR-properties or even characterise topological properties via IR-properties.

The next section deals with the separation properties T_0, T_1, and T_2 of the topologies τ, τ'', and τ'. We give a characterization of T_0 for τ, and of T_0, T_1, T_2, for τ'' and τ' in terms of the IR-property introduced above: the separation of the points of DS by the similarity functions $\mathrm{sim}(., Q)$, $Q \in QS$.

The third section deals with the question: which spaces (DS, τ), (DS, τ''), (DS, τ') are proximity spaces? These are spaces where there is a notion of "closeness", important in IR. We will show that τ'' and τ' are fine enough to be proximity spaces but that τ is not, in general.

The paper closes with the study of connectivity of the topologies τ, τ'', and τ'. A characterization of connectivity in terms of the availability of the NOT operator in QS (i.e., NOT $Q \in QS$ if $Q \in QS$) is given. This is another example of a topological characterization of certain IR-properties (or tools).

2. CHARACTERIZATION OF THE SEPARATION PROPERTIES T_0, T_1, AND T_2 FOR THE TOPOLOGIES τ, τ'', τ' ON DS

Recall (see, e.g., [3,7]) that a topological space (X, τ) is called a T_0-space if, whenever $x, y \in X$, $x \neq y$, there exists a neighborhood of x or of y not containing the other point. It is called a T_1-space if there exist neighborhoods of x and of y not containing the other point. Finally, it is called a T_2-space (or a Hausdorff space) if, whenever $x, y \in X$, $x \neq y$, there is a neighborhood U of x, and a neighborhood V of y such that $U \cap V = \phi$. Obviously, $T_2 \Rightarrow T_1 \Rightarrow T_0$ but not conversely. In fact, this will also be seen in this paper. For $x \in X$, let us denote $V_\tau(x)$ (or simply $V(x)$ if no confusion can arise), the set of neighborhoods of x in the topological space (X, τ). We will now investigate (DS, τ), (DS, τ''), and (DS, τ') w.r.t. these separation properties.

The stronger the separation property a space has, the finer are the possibilities of retrieval since one can make better distinction between documents. Indeed, in the above definitions of the separation properties T_0, T_1, and T_2, one can equally work with open sets (i.e., sets in the topology) as a substitute for the neighborhoods of the points. Now, as proved in [3] and repeated in the Introduction, these sets represent general Boolean queries in the sense that any Boolean query is represented (through its retrievals) by open sets and that any open set is a retrieval of a certain Boolean query. Hence, in case some separation property (T_0, T_1 or even stronger: T_2) is available there exist (Boolean) techniques to limit the search to certain documents, i.e., finer searches are possible.

We start with the easiest results.

THEOREM 2.1. *Let (DS, QS, sim) be any IR-model. For the similarity topology τ'', the following properties are equivalent.*

(i) *τ'' is T_0.*
(ii) *τ'' is T_1.*
(iii) *τ'' is T_2.*
(iv) *The retrieval model separates the points of DS.*

The proof is given in Appendix A. The same result is true for τ'. Also, this proof can be checked in Appendix A.

THEOREM 2.2. *Let (DS, QS, sim) be any IR-model. For the pseudometric topology τ', the following properties are equivalent.*

(i) *τ' is T_0.*
(ii) *τ' is T_1.*
(iii) *τ' is T_2.*
(iv) *τ' is a metric topology.*
(v) *The retrieval model separates the points of DS.*

The analogous result for τ (the retrieval topology) is not true. In fact, only the following result is true.

THEOREM 2.3. *Let (DS, QS, sim) be any IR-model. For the retrieval topology τ, the following properties are equivalent.*

(i) *τ is T_0.*
(ii) *The retrieval model separates the points of DS.*

The less trivial proof can also be checked in Appendix A. The T_1 analogue of Theorem 2.3 is as follows.

THEOREM 2.4. *Let (DS, QS, sim) be any IR-model. For τ, the following properties are equivalent.*

(i) *(DS, τ) is T_1.*
(ii) *For every $D_1, D_2 \in DS$, $D_1 \neq D_2$, there exist $Q, Q' \in QS$ such that*

$$\mathrm{sim}(D_2, Q) > \mathrm{sim}(D_1, Q), \qquad \mathrm{sim}(D_2, Q') < \mathrm{sim}(D_1, Q').$$

The proof is very similar to the one of Theorem 2.3 and hence is omitted.

Note in Case $DS = QS$

Although it is not necessary for the above to have that $DS = QS$, it is an important and interesting case in practice! If this is so, then (b) $\Rightarrow$ (a) $\Rightarrow$ (ii) in Theorem 2.3:

(a) $\forall\, D_1, D_2 \in DS$, $D_1 \neq D_2$,
$\mathrm{sim}(D_1, D_2) < \max(\mathrm{sim}(D_1, D_1), \mathrm{sim}(D_2, D_2))$,
(b) $\mathrm{sim}(D_1, D_2) = \mathrm{sim}(D_1, D_1) = \mathrm{sim}(D_2, D_2)$ implies $D_1 = D_2$.

Furthermore, (ii) of Theorem 2.4 is satisfied if

$$\mathrm{sim}(D_1, D_2) \notin [\mathrm{sim}(D_1, D_1), \mathrm{sim}(D_2, D_2)],$$

for every $D_1, D_2 \in DS$, $D_1 \neq D_2$. We leave the easy proofs to the reader.

Note in Case DS is Finite

Note that if DS is finite and T_1 (for any topology), then this topology is discrete, i.e., $\wp(DS)$. This follows from the fact that every singleton is closed.

It is already clear from Theorems 2.1–2.4 that the retrieval topology τ plays a special role in the sense that for τ the separation properties T_0, T_1, and T_2 seem to be different (in contrast to the cases of τ'' and τ'). We now present concrete evidence for this. In [3], we already gave examples of spaces (DS, τ) that are T_0 but not T_1. Hence (because of the above results), $\tau \neq \tau''$ also. Now, τ can be T_1 without being equal to τ'' as the next example shows.

EXAMPLE 2.5. There exist an IR-model (DS, QS, sim) such that (DS, τ) is T_1 and not T_2. Hence, also $\tau \neq \tau''$.

Proof. Note, that our example must be one for which DS is infinite, since otherwise (DS, τ) will be a finite T_1-space, hence discrete ($\tau = \wp(DS)$), and hence $\tau = \tau'' = \tau'$ (and T_2). Take $DS = QS = \mathbb{N}$. For all, $p, n \in \mathbb{N}$, define

$$\mathrm{sim}(n,p) = \max\left(\frac{n}{p}, \frac{p}{n}\right).$$

One has that $\forall\, n \in DS$:

$$DS \setminus \{n\} = R(n,1) \in \tau,$$

since $\mathrm{sim}(n,p) \geq 1$ and $\mathrm{sim}(n,p) = 1 \Leftrightarrow n = p$. Hence (DS, τ) is T_1. By Theorem 2.3, we hence know that the IR-model (DS, QS, sim) separates the points of DS. Hence (Theorem 2.1), (DS, τ'') is T_2. But τ is not T_2 because $\forall\, n, n' \in DS$, $n \neq n'$, $\forall\, U \in V_\tau(n)$, $\forall\, V \in V_\tau(n')$, $\exists\, a \in \mathbb{N}$ such that

$$\{m \in \mathbb{N} \parallel m \geq a\} \subset U \cap V.$$

Indeed, $\forall\, n, n' \in DS$, $\forall\, r, r' > 0$:

$$R(n,r) \cap R(n',r') \supset]\max(nr, n'r'), +\infty[\cap \mathbb{N}.$$

Hence $\tau \neq \tau''$. $\blacksquare$

Notes.

1. If we take $\mathrm{sim}(n,p) = n/p$, then τ is T_0 and not T_1 (but in [3], we had already such an example).
2. τ can even be T_2 and still $\tau \neq \tau''$. We can only show this in the next section (Example 3.3).

Interpretation of the Separation Properties T_0, T_1, T_2 in Terms of Boolean Queries

Let us focus on τ'' and T_2. If (DS, τ'') is T_2, i.e., when the retrieval model separates the points of DS (Theorem 2.1), then we have that, whenever $D_1, D_2 \in DS$, $D_1 \neq D_2$, there exist $U, V \in \tau''$, such that $D_1 \in U$ and $D_2 \in V$ and such that $U \cap V = \phi$. Since the sets in

$$\left\{ \bigcap_{i=1}^{n} U\left(Q_i, r_1^{(i)}, r_2^{(i)}\right) \parallel n \in \mathbb{N},\ Q_i \in QS,\ r_1^{(i)} < r_2^{(i)} \right\}$$

form a basis for τ'', it follows that (cf. [8, p. 38, 5.1.]) (equivalently) there exist $Q_1, \dots, Q_n$, $Q'_1, \dots, Q'_m \in QS$, $r_1^{(i)} < r_2^{(i)}$, $r_1'^{(j)} < r_2'^{(j)}$ such that

$$D_1 \in \bigcap_{i=1}^{n} U\left(Q_i, r_1^{(i)}, r_2^{(i)}\right), \qquad D_2 \in \bigcap_{j=1}^{m} U\left(Q'_j, r_1'^{(j)}, r_2'^{(j)}\right),$$

and such that

$$\left(\bigcap_{i=1}^{n} U\left(Q_i, r_1^{(i)}, r_2^{(i)}\right)\right) \cap \left(\bigcap_{j=1}^{m} U\left(Q'_j, r_1'^{(j)}, r_2'^{(j)}\right)\right) = \phi.$$

By definition, this is equivalent with the property that different documents imply the existence of two disjoint retrievals of Boolean AND queries such that each document belongs to one of these retrievals. This shows the degree of "fine-tuning" (as far as retrieval is concerned) that is possible in such spaces.

Let us give another example: the T_1-property in (DS, τ). First of all, the definition of T_1 gives: $\forall\, D_1, D_2 \in DS$, $D_1 \neq D_2$, $\exists\, U, V \in \tau$ such that $D_1 \in U$, $D_2 \notin U$, $D_1 \notin V$, $D_2 \in V$. Since the sets in

$$\left\{ \bigcap_{i=1}^{n} R(Q_i, r_i) \parallel Q_i \in QS, r_i \in \mathbb{R} \right\},$$

form a basis for τ, it follows that any of these documents can be excluded from a retrieval of a Boolean AND query that contains the other document: $\exists Q_1, \dots, Q_n \in QS$, $\exists r_1, \dots, r_n \in \mathbb{R}$ such that (e.g., for D_1)

$$D_1 \in \bigcap_{i=1}^{n} R(Q_i, r_i) \subset U \quad \text{and} \quad D_2 \notin \bigcap_{i=1}^{n} R(Q_i, r_i).$$

One can also look at T_1 in the following way: the T_1-property (of any topology) is equivalent with: every singleton is closed (cf. [8]). If DS is finite then any subset of DS is open for this topology (τ or $\tau'' = \tau'$). Hence, by definition of τ (analogously for $\tau'' = \tau'$), and by [3] (see also the argument around formula (9)), any subset of DS has the form

$$\bigcup_{j=1}^{m} \bigcap_{i=1}^{n_j} R(Q_{ij}, r_{ij}).$$

Here $Q_{ij} \in QS$, $r_{ij} \in \mathbb{R}$, $i = 1, \dots, n_j$, $j = 1, \dots, m$. We have proven the following result.

THEOREM 2.6. *Let (DS, τ) or (DS, τ'') be a T_1-space and let DS be finite. Then any subset of DS can be retrieved by using a Boolean query.*

Similar, equivalent, interpretations can be given for the other separation properties in conjunction with τ, τ'', or τ'.

We now present some examples to illustrate T_0, T_1, and T_2 in IR-models. For more information on these examples, see [3].

EXAMPLES 2.7.

EXAMPLE 2.7.1. THE VECTOR SPACE MODEL OF [3]. This example is as follows: $DS = QS = I^n$, ($I = [0,1]$), $\text{sim}(D,Q) = \langle D, Q \rangle$, $\forall D \in DS$, $\forall Q \in QS$, where $\langle D, Q \rangle$ denotes the inproduct between the vectors D and Q. In [3], it was shown that $\tau \neq \tau'' = \tau' = \mathcal{E}$, the Euclidean topology. It is clear that $\langle .,. \rangle$ separates the points of DS, and hence, τ is T_0 (and, of course, $\tau' = \tau''$ is T_2). τ is not T_1 [3].

EXAMPLE 2.7.2. THE VECTOR SPACE MODEL OF [1]. Here we have $DS = QS = I^n \setminus \{0\}$ and $QS^* = DS^* = DS/\Re$, the quotient space w.r.t. the relation $D_1 \Re D_2$ iff D_1 and D_2 are situated on the same half line through 0. For both DS and DS^*, one uses

$$\text{sim}(D,Q) = \text{sim}(D^*, Q^*) = \cos \widehat{(D,Q)},$$

the cosinus of the angle between the vectors $D, Q \in DS$ (or the half lines $D^*, Q^* \in DS^*$). We have that $\tau = \tau'' = \tau'$ on DS as well as on DS^*. However, points are separated in DS^*, but not in DS, so that DS^* is T_2, but DS is not even T_0 for any of the topologies $\tau = \tau'' = \tau'$.

EXAMPLE 2.7.3.

$$DS = QS = \{D_1, \dots, D_n\}, \qquad \text{sim}(D_i, D_j) = \frac{i+j}{2\max(i,j)}, \tag{10}$$

gives $\tau = \tau'' = \tau' = \wp(DS)$, the discrete topology. Hence, they are T_2. The same is true for the next two examples.

EXAMPLE 2.7.4.

$$DS = QS = \{D_1, \dots, D_n\}, \qquad \text{sim}(D_i, D_j) = \frac{2}{\pi} \operatorname{Arctan} \left(\frac{1}{|i-j|} \right). \tag{11}$$

EXAMPLE 2.7.5.

$$DS = QS: \text{ any sets, } \text{sim}(D,Q) \begin{cases} = 1, & D = Q, \\ = 0, & D \neq Q. \end{cases} \tag{12}$$

Example 2.7.6. An Example of a Space (DS, τ) that is not T_0. Take

$$DS = \{a, b, c, d\}, \qquad QS = \{e\},$$

and define

$$\mathrm{sim}(a, e) = \frac{1}{2} = \mathrm{sim}(d, e), \qquad \mathrm{sim}(b, e) = 1, \qquad \mathrm{sim}(c, e) = 0.$$

So the model does not separate the documents a and d. Hence, (DS, τ) is not T_0. In fact,

$$\tau = \{\phi, DS, \{a, b, d\}, \{b\}\} \quad \text{and} \quad \forall U \in V(a), \qquad \forall V \in V(d) : U \cap V \supset \{a, b, d\} \neq \phi.$$

These results show the real difference between the topologies τ, τ'', and τ'. Especially τ is shown to be the "rougher" one. Not only is it the coarsest topology of the three, it is *only* T_0 under the reasonable condition that the IR-model separates the points. Under this condition, τ'' as well as τ' are even T_2. This shows a big difference from the IR-point of view, a difference which is not revealed by just showing that the inclusions $\tau \subset \tau'' \subset \tau'$ can be strict.

The condition that the IR-model separates the points is "even more than reasonable". Indeed, if it is not the case, the different documents for which all sim$(., Q)$-values ($Q \in QS$) are equal, are not distinguishable from the IR-point of view. Hence, an equivalence relation (and subsequent quotient space) can be defined so that the new "points" are indeed separated now and the documents in the equivalence classes are considered as the same. In these cases, τ'' and τ' are always T_2 but τ is not. This shows the low separation capacity of τ for documents of DS.

3. PROXIMITY ASPECTS OF (DS, τ), (DS, τ''), AND (DS, τ')

Pseudo-metric spaces such as (DS, τ') bear in their proper definition the notion of proximity, by using the pseudo-metric (in the connection of (DS, τ'), we will always work with d' of formula (5) although other formulae are possible).

A proximity $\wp$ is defined between two subsets A, B of a set X and satisfies (by definition) the following properties (cf. [8]):

(P1) $A \wp B \iff B \wp A$,
(P2) $\{x\} \wp \{x\}, \forall x \in X$,
(P3) $A \wp (B \cup C) \iff A \wp B \vee A \wp C$,
(P4) $\phi \bar{\wp} X$,
(P5) $A \bar{\wp} B \Rightarrow \exists P, Q \subset X, P \cap Q = \phi$ such that $A \bar{\wp} P^c$ and $B \bar{\wp} Q^c$.

Here $\bar{\wp}$ means the negation of $\wp$ and P^c the X-complement $X \setminus P$.

$\wp$ defines the notion "to be near to each other" (or not, via $\bar{\wp}$). $(X, \wp)$ is called a proximity space. Every pseudo-metric space (X, d) is a proximity space as follows: define, for $A, B \subset X$, $A \wp B$ iff $d(A, B) = 0$. Here $d(A, B)$ is defined as

$$d(A, B) = \inf\{d(x, y) \parallel x \in A, \ y \in B\}. \tag{13}$$

From the above definitions, it is clear that proximity is a desirable property to have for an IR-model. Of course, (pseudo-)metrizability is even better, since then one can really measure distances between (sets of) documents. However, this cannot always be accomplished. As we will see in this section, (DS, τ) and (DS, τ'') are not always metrizable if the number of elementary requests (i.e., the cardinality of QS) is too high. In these cases, the notion of proximity is useful and it is worth investigating what spaces (DS, τ), (DS, τ''), or (DS, τ') are proximity spaces.

Of course, since τ' is pseudo-metrizable it is always a proximity space. What about τ and τ''? We first recall a few notions that are needed in the sequel.

Definition 3.1. A topological space (X, τ) is called completely regular if for every closed $F \subset X$ and $x \in X \backslash F$, there is a τ-continuous function $f : X \to [0, 1]$ such that $f(x) = 0$ and $f(F) = \{1\}$.

It is called regular if, whenever $F \subset X$ is closed and $x \in X \backslash F$, there are $U, V \in \tau$, $U \cap V = \phi$ such that $x \in U$ and $F \subset V$. Of course, complete regularity implies regularity (use $U = f^{-1}([0, 1/2[)$ and $V = f^{-1}(]1/2, 1])$ in the definition of complete regularity). Define a T_3-space to be a regular T_0-space and a $T_{3\,1/2}$-space to be a completely regular T_0-space. We have that $T_{3\,1/2} \Rightarrow T_3 \Rightarrow T_2$ obviously and it can be shown that the topological spaces that are proximity spaces are precisely the completely regular spaces.

NOTE. The notation $T_{3\,1/2}$ is a "joke" of topologists to denote a separation property that is logical situated between T_3 (introduced here) and T_4 (not introduced here). Some authors even use T_π instead of $T_{3\,1/2}$ (see [4, p. 167]).

The above definition gives another argument for the fact that finer searches are possible in proximity spaces.

We have the following result.

THEOREM 3.2. *Let the IR-system* (DS, QS, sim) *be such that the points of DS are separated (see the introductory section). Then*

(i) (DS, τ'') *is a proximity space,*

(ii) (DS, τ) *is not always a proximity space.*

The proof is given in Appendix B. An example showing that (DS, τ) is not always a proximity space (even when the points of DS are separated) is given by [3] (see also Example 2.7.1 here).

NOTE. From the proof of Theorem 3.2(i), it is clear that (DS, τ''), although it is a proximity space, it is not always pseudo-metrizable. Indeed, it is a subspace of $\prod_{Q \in QS} \mathbb{R}_Q (\mathbb{R}_Q = \mathbb{R}, \forall Q \in QS)$ and the latter one is only (pseudo-)metrizable if the cardinality of QS is less than or equal to that of $\mathbb{N}$. This is certainly not the case in the example given: here $QS = I^n$ which has the cardinality of the continuum. It would be good to have a characterization of the (pseudo-)metrizability of (DS, τ'') (and also of (DS, τ)).

A new example of a (DS, τ) that is not a proximity space now follows. This example, however, is also an example of a retrieval topology τ that is T_2, not T_3 and $\tau \neq \tau''$.

EXAMPLE 3.3. There exists an IR-model (DS, QS, sim) such that (DS, τ) is T_2 and not T_3. Hence, $\tau \neq \tau''$ in this case.

PROOF. (Based on [8, p. 92].) Take $DS = QS = \mathbb{R}^+$. $\forall n \in \mathbb{N}$, define

$$\mathrm{sim}\left(0, \frac{1}{n}\right) = 0,$$

$\forall p \in \mathbb{R}^+$ and $1/p \notin \mathbb{N}$:

$$\mathrm{sim}(0, p) = \frac{1}{1+p},$$

$\forall p, p' \in \mathbb{R}^+$, $p, p' \neq 0$:

$$\mathrm{sim}(p, p') = \frac{1}{1 + |p - p'|}.$$

One readily verifies that $\forall r < 1$,

$$R(0, r) = \left[0, \frac{1}{r} - 1\right[\setminus \left\{\frac{1}{n} \,\|\, n \in \mathbb{N}\right\},$$

and $\forall p \in \mathbb{R}^+ \setminus \{0\}$, $\forall r < 1$:

$$R(p, r) = \left]p - \frac{1}{r} + 1, p + \frac{1}{r} - 1\right[\cap \mathbb{R}^+.$$

Based on [8, p. 92, (14.2)], we see that (DS, τ) is a T_2-space but not a T_3-space (0 and the closed set $\{1/n \,\|\, n \in \mathbb{N}\}$ cannot be separated). Since τ is T_2, hence T_0, we have by Theorem 2.3 that the IR-model separates the points of DS. Hence, by Theorems 2.1 and 3.2, τ'' is a T_0 proximity space, hence it is $T_{3\,1/2}$, and hence T_3. So $\tau \neq \tau''$. ∎

4. CONNECTIVITY OF THE SPACES (DS, τ), (DS, τ''), (DS, τ') AND ITS RELATION TO THE BOOLEAN NOT-OPERATOR

In this last section, we deal with the topological notion of connectivity of the different topologies on DS. A topological space is said to be connected if it is not disconnected. A topological space (X, τ) is said to be disconnected if there exist $U, V \in \tau$ such that $\phi \neq U$, $\phi \neq V$, $U \cap V = \phi$, and $U \cup V = X$.

The study of (dis)connectedness in the topological spaces (DS, τ), (DS, τ''), and (DS, τ') has its relevance to IR in the following way: disconnected parts of DS divide the document space in "predefined" subsets that will mark a separation in the retrieval results: most commonly if documents of one part are retrieved, the ones of the other parts are not (using the same type of query). It is, therefore, not so surprising (intuitively) that the Boolean NOT-operator is involved here. As introduced in [3] and repeated in the Introduction, the general Boolean query

$$\underset{j=1}{\overset{m}{\mathrm{OR}}} \left(\underset{i=1}{\overset{n}{\mathrm{AND}}}\, Q_{ij} \right), \tag{14}$$

was defined through its set of retrievals (e.g., in τ-one can define it also in τ'')

$$\operatorname{ret}\left(\underset{j=1}{\overset{m}{\mathrm{OR}}} \left(\underset{i=1}{\overset{n}{\mathrm{AND}}}\, Q_{ij} \right) \right) = \left\{ \bigcup_{j=1}^{m} \left(\bigcap_{i=1}^{n} R(Q_{ij}, r_{ij}) \right) \parallel r_{ij} \in \mathbb{R} \right\}. \tag{15}$$

In the same way, we will now define the Boolean operator NOT. Note that (14) does not necessarily belong to QS. The same remark will go for NOT.

4.1 The Boolean NOT-Operator and First Topological Properties

Let (DS, QS, sim) be any IR-system. Let $Q \in QS$ be any elementary query. Then we define NOT Q as a query (not necessarily belonging to QS) defined by the set of retrievals (for τ):

$$\operatorname{ret}(\text{NOT } Q) = \{R^c(Q, r) \parallel r \in \mathbb{R}\} \tag{16}$$

($R^c(Q, r) = DS \setminus R(Q, r)$, the complement of $R(Q, r)$). Only in case there exists a query $Q' \in QS$ such that

$$\{R(Q', r) \parallel r \in \mathbb{R}\} = \{R^c(Q, r) \parallel r \in \mathbb{R}\}, \tag{17}$$

we can identify Q' and NOT Q, and hence, we can then consider NOT $Q \in DS$.

EXAMPLE.

$$QS_1 = \{\text{cat, dog, shoe, horse}\},$$
$$QS_2 = \{\text{cat, dog, honest, dishonest}\}.$$

We have $Q \in QS_1 \Rightarrow$ NOT $Q \notin QS_1$. For $Q =$ "honest" $\in QS_2$, we have that NOT $Q \in QS_2$.

NOTE.

(1) (17) does *not* imply that $R(Q', r) = R^c(Q, r)$,

(2) By using $U(Q, r_1, r_2)$ instead of $R(Q, r)$, we can define NOT Q w.r.t. τ'. To be exact, we should have indicated in the notation of NOT Q with which topology we are working. We did not do so, however, since it will be clear from the context and for the sake of simplicity.

Let $Q \in QS$. Let us call Q trivial (w.r.t. τ) if

$$\{R(Q,r) \| r \in \mathbb{R}\} \subset \{\phi, DS\}. \tag{18}$$

The same definition is possible w.r.t. τ'': we call $Q \in QS$ trivial w.r.t. τ'' if

$$\{U(Q,r_1,r_2) \| r_1 < r_2\} \subset \{\phi, DS\}. \tag{19}$$

Both notions, however, are equivalent as the next proposition shows.

PROPOSITION 4.1.1. *Let (DS, QS, sim) be any IR-model and $Q \in QS$ arbitrary. Then Q is trivial w.r.t. τ iff it is trivial w.r.t. τ''.*

For the proof, we refer the reader to Appendix C. From now on, we simply call Q trivial without referring to τ or τ''. Note that the notion of NOT Q depends on τ or τ'' as is clear from the next proposition (for τ) and counterexample (for τ'').

PROPOSITION 4.1.2. *If $Q \in QS$ is nontrivial, then $Q \neq$ NOT Q (w.r.t. τ).*

PROOF. Suppose $Q =$ NOT Q. Then

$$\{R(Q,r) \| r \in \mathbb{R}\} = \{R^c(Q,r) \| r \in \mathbb{R}\} \not\subset \{\phi, DS\}.$$

Let then $R(Q,r_0)$ be such that $R(Q,r_0) \neq \phi$, $R(Q,r_0) \neq DS$. Hence $D, E \in DS$ exist such that $D \in R(Q,r_0)$, $E \notin R(Q,r_0)$. So

$$\mathrm{sim}(D,Q) > r_0 \geq \mathrm{sim}(E,Q). \tag{20}$$

By the above equality, there is a $r_1 \in \mathbb{R}$ such that $D \in R^c(Q,r_1)$, $E \notin R^c(Q,r_1)$. Hence, $D \notin R(Q,r_1)$ and $E \in R(Q,r_1)$. So

$$\mathrm{sim}(E,Q) > r_1 \geq \mathrm{sim}(D,Q). \tag{21}$$

(20) contradicts (21). ∎

COUNTEREXAMPLE 4.1.3. The above proposition is not true for τ''. Indeed: take $DS = \{a,b\}$, $QS = \{c\}$, $\mathrm{sim}(a,c) = 1/4$, $\mathrm{sim}(b,c) = 1/2$. Hence

$$\{U(c,r_1,r_2) \| r_1 < r_2\} = \{\phi, DS, \{a\}, \{b\}\} = \wp(DS).$$

Hence, c is not trivial and by the above

$$\{U^c(c,r_1,r_1) \| r_1 < r_2\} = \{U(c,r_1,r_2) \| r_1 < r_2\},$$

showing that $c =$ NOT c (w.r.t. τ'').

COROLLARY 4.1.4. *There exist $Q \in QS$ such that NOT Q (w.r.t. τ) $\neq$ NOT Q (w.r.t. τ'').*

This cannot lead to any confusion, since we will make it very clear whether we are working in a "threshold" environment (τ) or a "close match" environment (τ'').

PROPOSITION 4.1.5. *If $QS = \{Q\}$ with Q nontrivial (w.r.t. τ). Then (DS, τ) is connected.*

PROOF. For any open sets $U, V \in \tau$, $U, V \neq \phi$, $\exists\, n \in \mathbb{N}$, $\exists\, Q_i \in QS$, $\exists\, r_i \in \mathbb{R}$, $i = 1, \dots, n$ such that

$$\phi \neq \bigcap_{i=1}^{n} R(Q_i, r_i) \subset U,$$

$\exists m \in \mathbb{N}$, $\exists Q'_j \in QS$, $\exists r'_j \in \mathbb{R}$, $j = 1, \ldots, m$ such that

$$\phi \neq \bigcap_{j=1}^{m} R(Q'_j, r'_j) \subset V.$$

Since all $Q_i, Q'_j = Q$, we have, if $r''_0 = \max_{i=1,\ldots,n} r_i$, $r''_1 = \max_{j=1,\ldots,m} r'_j$,

$$\phi \neq R(Q, r''_0) \subset U, \qquad \phi \neq R(Q, r''_1) \subset V.$$

Let $r = \max(r''_0, r''_1)$. Then

$$\phi \neq R(Q, r) \subset U \cap V.$$

So (DS, τ) cannot be disconnected. ∎

We will give an example in the sequel (see Counterexample 4.2.3) showing that (DS, τ) can be disconnected as soon as QS has two nontrivial elements.

Proposition 4.1.5 is not valid for τ'': take $DS = \{a, b, c, d\}$, $QS = \{e\}$, $\text{sim}(a, e) = \text{sim}(d, e) = 1/2$, $\text{sim}(b, e) = 1$, $\text{sim}(c, e) = 0$. Then e is nontrivial w.r.t. τ'' (this is clear) and

$$\tau'' = \{\{a, b, c\}, \{b\}, \{a, c, d\}, \{c\}, DS, \phi\}.$$

So (DS, τ'') is disconnected (by $\{b\}$ and $\{a, c, d\}$). This is also an example of a space such that (DS, τ) is connected, but (DS, τ'') is not.

For general IR-models (DS, QS, sim), we can ask the question: when is (DS, τ) or (DS, τ'') disconnected? This will be studied in the next section, where the NOT-operator will be the key element.

4.2. Characterization of Connectivity Using the NOT-Operator

The next proposition shows that, with our notion of the NOT-operator, we are heading in the right direction when we want to characterize connectivity.

PROPOSITION 4.2.1. *Let (DS, QS, sim) be any IR-model. If $Q \in QS$ is a nontrivial query such that (w.r.t. τ) NOT $Q \in QS$, then (DS, τ) is disconnected.*

PROOF. So $Q \in QS$ is such that

$$\{R(Q, r) \parallel r > 0\} \not\subset \{\phi, DS\}.$$

Hence,

$$\{R^c(Q, r) \parallel r > 0\} \not\subset \{\phi, DS\}.$$

Since NOT $Q \in DS$, there exists $r_1, r_2 \in \mathbb{R}$ such that

$$\phi \neq R^c(Q, r_1) = R(\text{NOT } Q, r_2) = DS.$$

Hence, $\{R(Q, r_1), R(\text{NOT } Q, r_2)\}$ forms a nontrivial τ-open disconnection of DS. Hence (DS, τ) is disconnected. ∎

The same result is true for τ''.

PROPOSITION 4.2.2. *Let (DS, QS, sim) be any IR-model. If $Q \in QS$ is a nontrivial query such that (w.r.t. τ'') NOT $Q \in QS$, then (DS, τ'') is disconnected.*

The proof is similar to that of Proposition 4.2.1. These results do not constitute a characterization of (dis)connectivity since the converses of the above theorems are not true, as the next example shows.

Counterexample 4.2.3. Let $DS = \{a, b, c, d\}$, $QS = \{e, f\}$ and

$$\begin{aligned} \mathrm{sim}(a,e) &= \frac{1}{2}, & \mathrm{sim}(a,f) &= \frac{1}{2}, \\ \mathrm{sim}(b,e) &= \frac{3}{4}, & \mathrm{sim}(b,f) &= \frac{3}{4}, \\ \mathrm{sim}(c,e) &= \frac{1}{4}, & \mathrm{sim}(c,f) &= 1, \\ \mathrm{sim}(d,e) &= \frac{1}{2}, & \mathrm{sim}(d,f) &= \frac{1}{2}. \end{aligned}$$

Then

$$\begin{aligned} \tau &= \{DS, \phi, \{a,b,d\}, \{b\}, \{b,c\}, \{c\}\}, \\ \tau'' &= \{DS, \phi, \{a,b,d\}, \{b\}, \{b,c\}, \{c\}, \{a,d\}, \{a,c,d\}\}, \\ \{R(e,r) \parallel r \in \mathbb{R}\} &= \{\phi, DS, \{a,b,d\}, \{b\}\}, \\ \{R(f,r) \parallel r \in \mathbb{R}\} &= \{\phi, DS, \{b,c\}, \{c\}\}, \\ \{U(e,r_1,r_2) \parallel r_1 < r_2\} &= \{\phi, DS, \{c\}, \{a,c,d\}, \{a,d\}, \{a,b,d\}, \{b\}\}, \\ \{U(f,r_1,r_2) \parallel r_1 < r_2\} &= \{\phi, DS, \{a,d\}, \{a,b,d\}, \{b\}, \{b,c\}, \{c\}\}. \end{aligned}$$

It is hence clear that NOT $Q \notin QS$ for $Q \in QS$ for τ as well as τ'', yet τ and τ'' are disconnected: τ by $\{\{a,b,d\}, \{c\}\}$ and τ'' by the same set and also by $\{\{a,c,d\}, \{b\}\}$ and even by $\{\{a,d\}, \{b,c\}\}$.

Incidentally, this example also shows, as promised, that (DS, τ) can be disconnected as soon as QS has more than one element.

Note also the fact that τ'' is "much more" disconnected than τ, a logical fact. In this context, it is interesting to look at the connected components of the spaces (DS, τ) and (DS, τ''). These are the maximally connected subsets of DS (for τ, respectively, τ'') (see, e.g., [8, p. 194]). (DS, τ) has only two components: $DS = \{a,b,d\} \dot{\cup} \{c\}$ ($\dot{\cup}$ denotes disjoint union) for τ and three for $\tau'' : DS = \{b\} \dot{\cup} \{c\} \dot{\cup} \{a,d\}$.

The next theorem yields a characterization of connectivity in terms of the NOT-operator.

Theorem 4.2.4. *Let (DS, QS, sim) be any IR-model. For (DS, τ), the following assertions are equivalent.*

(i) *(DS, τ) is disconnected.*
(ii) *There exist arrays $(Q_{ij}), (Q'_{\ell k})$, $(j \in J,\ i \in \{1, \dots, n_j\},\ n_j \in \mathbb{N},\ k \in K,\ \ell \in \{1, \dots, p_k\}$, $p_k \in \mathbb{N})$ in QS and there exist retrievals (w.r.t. τ)*

$$A_{ij} \in \mathrm{ret}(Q_{ij}), \tag{22}$$

$$A'_{\ell k} \in \mathrm{ret}(\text{NOT } Q'_{\ell k}), \tag{23}$$

such that

$$\phi \neq \bigcup_{j \in J} \bigcap_{i=1}^{n_j} A_{ij} = \bigcap_{k \in K} \bigcup_{\ell=1}^{p_k} A'_{\ell k} \neq DS. \tag{24}$$

The proof is given in Appendix D.

Note that (ii) is satisfied in the following cases:

(a) $\exists\, Q, Q' \in QS$ such that $\mathrm{ret}(Q) \cap \mathrm{ret}(\text{NOT } Q') \neq \{\phi, DS\}$,
(b) $\exists\, Q \in QS$, nontrivial, such that $\mathrm{ret}(Q) = \mathrm{ret}(\text{NOT } Q)$,
(c) $\exists\, Q \in QS$, nontrivial, such that NOT $Q \in QS$ (cfr. Proposition 4.2.1).

Indeed, for (a) take J and K as singletons, $n_j = p_k = 1$, (b) ⇒ (a). For (c), take $Q' =$ NOT Q in (a).

THEOREM 4.2.5. *Theorem 4.2.4 with τ replaced by τ'' (and of course using* $\mathrm{ret}_{\tau''}$ *instead of* ret_τ*) is also valid. The proof is the same as the one of Theorem 4.2.4, and hence, is omitted.*

For finite document spaces DS, Theorems 4.2.4 and 4.2.5 have the following, rather surprising, consequence.

THEOREM 4.2.6. *Let (DS, QS, sim) be any IR-model. If DS is finite then the following assertions are equivalent (for τ, respectively, τ'').*

(i) *DS is connected.*

(ii) *There does not exist a Boolean retrieval (other than ϕ or DS) based on elementary queries (in QS) that is equal to a Boolean retrieval based on* NOT*s of elementary queries in QS.*

For the proof, we again refer the reader to Appendix D. Of course, a Boolean query based on NOTs of elementary queries in QS is defined in an analogous way as in the introduction: formulae (8) and (9) but now for NOT Q_{ij}, instead of Q_{ij} and with $R(Q_{ij}, r_{ij})$ replaced by $R^c(Q_{ij}, r_{ij})$.

PROBLEM. It remains an interesting problem to determine the connected components of the spaces (DS, τ) and (DS, τ'').

REMARKS.

(1) Totally disconnected spaces are spaces in which the only connected subsets are the singletons. These spaces exist, namely, for any set X, take $\tau = \wp(X)$, the set of all subsets of X. In [3], several examples are given of spaces DS where $\tau = \tau' = \tau'' = \wp(DS)$, and hence, these IR-models are totally disconnected. The simplest of these models is the so-called discrete retrieval (see, e.g., Example 2.7.5): for any set DS, take $QS = DS$ and

$$\mathrm{sim}(D, Q)\begin{cases} = 1, & \text{if } D = Q, \\ = 0, & \text{if } D \neq Q. \end{cases}$$

(2) In the case of Theorem 4.2.6 (DS finite), we have that $\tau' = \tau''$, hence Theorem 4.2.6 characterizes connectivity of (DS, τ') as well! That $\tau' = \tau''$ follows from [3] as mentioned in the introduction here.

5. SUMMARY

In this paper, several topological properties of the spaces (DS, τ), (DS, τ''), and (DS, τ') are studied, for any IR-model (DS, QS, sim). Properties as T_0, T_1, and T_2 are characterized as well as determined whether or not these spaces are proximity spaces. Also characterizations of connectivity in terms of the Boolean NOT-operator are given.

In all these results, the relations between topological properties and IR-aspects are given.

APPENDIX A
PROOF OF THE SEPARATION PROPERTIES

THEOREM A.1. *Let (DS, QS, sim) be any IR-model. For the similarity topology τ'' the following properties are equivalent.*

(i) *τ'' is T_0.*

(ii) *τ'' is T_1.*

(iii) *τ'' is T_2.*

(iv) *The retrieval model separates the points of DS.*

PROOF. (iii) ⇒ (ii) ⇒ (i) is trivial and well known.

(iv) ⇒ (iii): Let $D_1, D_2 \in DS$, $D_1 \neq D_2$. By (iv), there is a $Q \in QS$ such that

$$\mathrm{sim}(D_1, Q) \neq \mathrm{sim}(D_2, Q).$$

We keep the full generality by assuming

$$\mathrm{sim}(D_1, Q) > \mathrm{sim}(D_2, Q).$$

Take

$$r = \frac{\mathrm{sim}(D_1, Q) + \mathrm{sim}(D_2, Q)}{2},$$

and define

$$\begin{aligned} A_1 &= R(Q, r) = \{D \in DS \,\|\, \mathrm{sim}(D, Q) > r\} \in \tau \subset \tau'', \\ A_2 &= \{D \in DS \,\|\, \mathrm{sim}(D, Q) < r\} \in \tau'' \end{aligned}$$

(A_2 does not always belong to τ!). Then $D_1 \in A_1$, $D_2 \in A_2$, and $A_1 \cap A_2 = \phi$. Hence τ'' is T_2.

(i) ⇒ (iv). Let $D_1, D_2 \in DS$, $D_1 \neq D_2$. Since (DS, τ'') is T_0 there exist a $U_1 \in \tau''$ such that $D_1 \in U_1$ and $D_2 \notin U_1$ OR there exists a $U_2 \in \tau''$ such that $D_2 \in U_2$ and $D_1 \notin U_2$. Let us suppose the first case. Since $U_1 \in \tau''$, there is an $n \in \mathbb{N}$, $Q_i \in QS$, $r_1^{(i)}, r_2^{(i)} \in \mathbb{R}$ $(i = 1, \dots, n)$ such that

$$D_1 \in \bigcap_{i=1}^{n} \mathrm{sim}^{-1}(., Q_i) \left( ]r_1^{(i)}, r_2^{(i)}[\right) \subset U_1.$$

So, $r_1^{(i)} < \mathrm{sim}(D_1, Q_i) < r_2^{(i)}$ for every $i = 1, \dots, n$. Since $D_2 \notin U_1$, we have that there exists a $i_0 \in \{1, \dots, n\}$ such that $\mathrm{sim}(D_2, Q_{i_0}) \notin]r_1^{(i_0)}, r_2^{(i_0)}[$. Hence $\mathrm{sim}(D_1, Q_{i_0}) \neq \mathrm{sim}(D_2, Q_{i_0})$. Hence the IR-model separates the points of DS. ∎

THEOREM A.2. *Let (DS, QS, sim) be any IR-model. For the pseudo-metric topology τ', the following properties are equivalent.*

(i) *τ' is T_0.*
(ii) *τ' is T_1.*
(iii) *τ' is T_2.*
(iv) *τ' is a metric topology.*
(v) *The retrieval model separates the points of DS.*

PROOF. (iv) ⇔ (iii) ⇒ (ii) ⇒ (i) is trivial and well known.

(v) ⇒ (iv). Let $D_1, D_2 \in DS$, $D_1 \neq D_2$ and let $Q \in QS$ be such that $\mathrm{sim}(D_1, Q) \neq \mathrm{sim}(D_2, Q)$. Hence,

$$d'(D_1, D_2) = \sup_{Q \in QS} \frac{|\,\mathrm{sim}(D_1, Q) - \mathrm{sim}(D_2, Q)|}{1 + |\,\mathrm{sim}(D_1, Q) - \mathrm{sim}(D_2, Q)|} > 0.$$

Hence, d' is a metric and hence τ' is a metric topology.

(i) ⇒ (v). Let $D_1, D_2 \in DS$, $D_1 \neq D_2$. Hence, there is an open d'-ball $B(D_1, \varepsilon)$ around D_1 such that $D_2 \notin B(D_1, \varepsilon)$, OR there is an open d'-ball $B(D_2, \varepsilon)$ around D_2 such that $D_1 \notin B(D_2, \varepsilon)$. In both cases, is $d'(D_1, D_2) > 0$, from which it easily follows that there exists a $Q \in QS$ such that

$$\mathrm{sim}(D_1, Q) \neq \mathrm{sim}(D_2, Q). \qquad ∎$$

THEOREM A.3. *Let (DS, QS, sim) be any IR-model. For the retrieval topology τ, the following properties are equivalent.*

(i) *τ is T_0.*
(ii) *The retrieval model separates the points of DS.*

PROOF. (ii) ⇒ (i). Let $D_1, D_2 \in DS$, $D_1 \neq D_2$. By (ii), there is a $Q \in QS$ such that $\text{sim}(D_1, Q) \neq \text{sim}(D_2, Q)$.

(a) Suppose that $\text{sim}(D_2, Q) < \text{sim}(D_1, Q)$. Let

$$r_0 = \frac{\text{sim}(D_1, Q) + \text{sim}(D_2, Q)}{2}.$$

Hence, $D_1 \in R(Q, r_0)$, $D_2 \notin R(Q, r_0)$, and $R(Q, r_0) \in \tau$. Hence (i) is valid.

(b) Suppose that $\text{sim}(D_1, Q) > \text{sim}(D_2, Q)$. The same argument as above, but with the indices 1 and 2 interchanged, yields (i).

(i) ⇒ (ii). For every $D_1, D_2 \in DS$, $D_1 \neq D_2$, let (by (i)) $V \in V_\tau(D_1)$ be such that $D_2 \notin V$ OR let $U \in V_\tau(D_2)$ be such that $D_1 \notin U$. Suppose we are in the first case. By the definition of τ, there exists $n \in \mathbb{N}$, $Q_i \in QS$, $r_i \in \mathbb{R}$ $(i = 1, \dots, n)$ such that

$$D_1 \in \bigcap_{i=1}^{n} R(Q_i, r_i) \subset V.$$

Since $D_2 \notin V$, there is a $i_0 \in \{1, \dots, n\}$ such that $D_2 \notin R(Q_{i_0}, r_{i_0})$. These results show that

$$\text{sim}(D_1, Q_{i_0}) > r_{i_0} \geq \text{sim}(D_2, Q_{i_0}),$$

and hence, that $\text{sim}(D_1, Q_{i_0}) \neq \text{sim}(D_2, Q_{i_0})$. The same argument applies to the second case. ∎

APPENDIX B
PROOF OF THE PROXIMITY PROPERTIES

THEOREM B.1. *Let the IR-system (DS, QS, sim) be such that the points of DS are separated. Then*

(i) *(DS, τ'') is a proximity space;*

(ii) *(DS, τ) is not always a proximity space.*

PROOF.

(i) By definition, τ'' = the weak topology generated by the functions $\text{sim}(., Q)$, $Q \in QS$ and these functions separate points of DS. By Theorem 8.12 in [8], p. 56 (DS, τ'') is homeomorphic to a subspace of

$$\prod_{Q \in QS} \mathbb{R}_Q,$$

where $\mathbb{R}_Q = \mathbb{R}$ for every $Q \in QS$, equipped with the product topology, via the mapping

$$D \in DS \to (\text{sim}(D, Q))_{Q \in QS}.$$

Now $\mathbb{R}$ is a completely regular (metric!) space and products of such spaces are completely regular ([8, p. 95]) and hence they are proximity spaces.

(ii) If (DS, τ) was always a proximity space, it would be a T_2 space since separation of the points of DS implies that (DS, τ) is T_0 (Theorem 2.3) by the properties in Section 3. This is shown not always to be true: see, e.g., Example 2.7.1: the vector space model of [3]: (DS, τ) is T_0, hence the IR-system separates the points of DS but it is not T_1, let alone T_2. ∎

PROBLEM. Characterize the T_2 property as well as the proximity property for the retrieval topology τ.

As a consequence of the above proof, we have the following proposition.

PROPOSITION. *Equip $\mathbb{R}$ with the topology generated by the set*

$$\{]r, +\infty[\,\|\, r \in \mathbb{R}\}.$$

Denote this topology by D. Hence,

$$D = \{\phi, \mathbb{R},]r, +\infty[, r \in \mathbb{R}\}.$$

Then

(i) *τ on DS is the coarsest topology making all the* $\text{sim}(.,Q) : DS \to \mathbb{R}, D$ $(Q \in QS)$ *continuous;*

(ii) *suppose that* (DS, QS, sim) *separates the points of DS. Then (DS, τ) is homeomorphic with a subspace of*

$$\prod_{Q \in QS} E_Q,$$

where $E_Q = (\mathbb{R}, D)$ for every $Q \in QS$.

PROOF.

(i) τ is generated by the sets

$$R(Q, r) = \{D \in DS \parallel \text{sim}(D,Q) > r\}.$$

Hence, it is the coarsest topology making all the functions $\text{sim}(.,Q) : DS \to (\mathbb{R}, D)$ continuous. This follows from the definition of D and the properties of the inverse relation.

(ii) Since the IR-model separates the points of DS, we can use [8, p. 56, Theorem 8.12] again, yielding that (DS, τ) is homeomorphic with a subspace of $\prod_{Q \in QS} E_Q$, where $E_Q = (\mathbb{R}, D)$ for all $Q \in QS$. ∎

COROLLARY. *The above results give a second proof of Theorem 2.3, (ii) ⇒ (i).*

Indeed, by (ii) of Theorem 2.3, (DS, τ) is homeomorphic with a subspace of $\prod_{Q \in QS} E_Q$. Since all $E_Q = (\mathbb{R}, D)$ are T_0 and since subspaces and products of T_0-spaces are T_0 ([8, p. 85]), the result follows.

APPENDIX C
PROOF OF PROPOSITION 4.1.1

PROPOSITION C.1. *Let* (DS, QS, sim) *be any IR-model and $Q \in QS$ arbitrary. Then Q is trivial w.r.t. τ iff it is trivial w.r.t. τ''.*

PROOF.

(i) Q trivial w.r.t. τ'' ⇒ Q trivial w.r.t. τ. So, $Q \in QS$ is such that

$$\{U(Q, r_1, r_2) \parallel r_1 < r_2\} \subset \{\phi, DS\}.$$

Let r_1 be fixed. We have that

$$R(Q, r_1) = \bigcup_{r > r_1} U(Q, r_1, r) \in \{\phi, DS\},$$

for every $r_1 \in \mathbb{R}$. Hence, Q is trivial w.r.t. τ.

(ii) Q trivial w.r.t. τ ⇒ Q trivial w.r.t. τ''. So, $Q \in QS$ is such that

$$\{R(Q, r) \parallel r \in \mathbb{R}\} \subset \{\phi, DS\}.$$

But $\forall r_1, r_2 \in \mathbb{R}$, $r_1 < r_2$:

$$\begin{aligned} U(Q, r_1, r_2) &= \{D \in DS \parallel r_1 < \text{sim}(D,Q) < r_2\} \\ &= \{D \in DS \parallel \text{sim}(D,Q) > r_1\} \setminus \{D \in DS \parallel \text{sim}(D,Q) \geq r_2\} \\ &= \{D \in DS \parallel \text{sim}(D,Q) > r_1\} \setminus \bigcap_{n=1}^{\infty} \left\{D \in DS \parallel \text{sim}(D,Q) > r_2 - \frac{1}{n}\right\} \\ &= R(Q, r_1) \setminus \bigcap_{n=1}^{\infty} R\left(Q, r_2 - \frac{1}{n}\right) \in \{\phi, DS\}. \end{aligned}$$

Hence, Q is trivial w.r.t. τ''. ∎

Note, that the countable intersection of τ-open sets is not necessarily open in τ. This is why this argument cannot be used to show that $U(Q, r_1, r_2)$ belongs to τ, in fact it is not, in general.

APPENDIX D
PROOF OF THE CHARACTERIZATION THEOREMS FOR CONNECTIVITY

THEOREM D.1. *Let (DS, QS, sim) be any IR-model. For (DS, τ) the following assertions are equivalent.*

(i) *(DS, τ) is disconnected.*

(ii) *There exist arrays $(Q_{ij}), (Q'_{\ell k})$ $(j \in J,\ i \in \{1, \dots, n_j\},\ n_j \in \mathbb{N},\ k \in K,\ \ell \in \{1, \dots, p_k\},\ p_k \in \mathbb{N})$ in QS and there exist retrievals*

$$A_{ij} \in \text{ret}(Q_{ij}), \tag{22}$$

$$A'_{\ell k} \in \text{ret}(\text{NOT } Q'_{\ell k}), \tag{23}$$

such that

$$\phi \neq \bigcup_{j \in J} \bigcap_{i=1}^{n_j} A_{ij} = \bigcap_{k \in K} \bigcup_{\ell=1}^{p_k} A'_{\ell k} \neq DS. \tag{24}$$

PROOF. (ii) $\Rightarrow$ (i). Let $Q_{ij}, Q'_{\ell k}$ be as given. Equations (22)–(24) imply $\forall\, i, j,\ \exists\, r_{ij} \in \mathbb{R}$ such that

$$A_{ij} = R(Q_{ij}, r_{ij}),$$

and $\forall\, \ell, k,\ \exists\, r'_{\ell k} \in \mathbb{R}$ such that

$$A'_{\ell k} = R^c(Q'_{\ell k}, r'_{\ell k}), \qquad \text{with}$$

$$\phi \neq \bigcup_{j \in J} \bigcap_{i=1}^{n_j} R(Q_{ij}, r_{ij}) = \bigcap_{k \in K} \bigcup_{\ell=1}^{p_k} R^c(Q'_{\ell k}, r'_{\ell k}) \neq DS.$$

Hence

$$\phi \neq \bigcup_{j \in J} \bigcap_{i=1}^{n_j} R(Q_{ij}, r_{ij}) = DS \setminus \bigcup_{k \in K} \bigcap_{\ell=1}^{p_k} R(Q'_{\ell k}, r'_{\ell k}) \neq DS.$$

Hence

$$G = \bigcup_{j \in J} \bigcap_{i=1}^{n_j} R(Q_{ij}, r_{ij}) \in \tau,$$

$$H = \bigcup_{k \in K} \bigcap_{\ell=1}^{p_k} R(Q'_{\ell k}, r'_{\ell k}) \in \tau,$$

forms a disconnection of (DS, τ).

(i) $\Rightarrow$ (ii). If (DS, τ) is disconnected then there exist $G, H \in \tau$, $\phi \neq G$, $DS \neq G$, $G \cap H = \phi$ such that $G \cup H = DS$. By definition of τ, there exist (Q_{ij}), $(Q'_{\ell k})$, r_{ij}, $r'_{\ell k} \in \mathbb{R}$, $j \in J$, $i = \{1, \dots, n_j\}$, $n_j \in \mathbb{N}$, $k \in K$, $\ell \in \{1, \dots, p_k\}$, $p_k \in \mathbb{N}$, such that

$$H = \bigcup_{j \in J} \bigcap_{i=1}^{n_j} R(Q_{ij}, r_{ij}),$$

$$G = \bigcup_{k \in K} \bigcap_{\ell=1}^{p_k} R(Q'_{\ell k}, r'_{\ell k}).$$

Since $H = G^c$, we have that

$$\phi \neq \bigcup_{j \in J} \bigcap_{i=1}^{n_j} R(Q_{ij}, r_{ij}) = \bigcap_{k \in K} \bigcup_{\ell=1}^{p_k} R^c(Q'_{\ell k}, r'_{\ell k}) \neq DS.$$

Obviously,

$$A_{ij} = R(Q_{ij}, r_{ij}) \in \mathrm{ret}(Q_{ij}),$$
$$A'_{\ell k} \in R^c(Q'_{\ell k}, r'_{\ell k}) \in \mathrm{ret}(\text{ NOT } Q'_{\ell k}). \qquad \blacksquare$$

THEOREM D.2. *Let (DS, QS, sim) be any IR-model. If DS is finite then the following assertions are equivalent (for τ, respectively, τ'').*

(i) *DS is connected.*

(ii) *There does not exist a Boolean retrieval (other than ϕ or DS) based on elementary queries (in QS) that is equal to a Boolean retrieval based on* NOT *s of elementary queries in QS.*

PROOF. The proof is, in essence, a negation of Theorems 4.2.4 and 4.2.5. For finite spaces DS, (24) looks like $(m, q \in \mathbb{N})$

$$\phi \neq \bigcup_{j=1}^{m} \bigcap_{i=1}^{n_j} A_{ij} = \bigcap_{k=1}^{q} \bigcup_{\ell=1}^{p_k} A'_{\ell k} \neq DS. \tag{25}$$

Now, the lemma below allows to write

$$\bigcap_{k=1}^{q} \bigcup_{\ell=1}^{p_k} A'_{\ell k} = \bigcup_{s=1}^{v} \bigcap_{r=1}^{u_s} A''_{rs},$$

where u_s, $v \in \mathbb{N}$ and where

$$\{A''_{rs} \parallel r = 1, \dots, u_s, s = 1, \dots, v\} = \{A'_{\ell k} \parallel \ell = 1, \dots, p_k, k = 1, \dots, q\}.$$

So, both sides of (25) are of the form

$$\bigcup_{\beta \in C_2} \bigcap_{\alpha \in C_1} A_{\alpha\beta},$$

where C_1 and C_2 are finite sets and where (for τ) on the left side of (24) sets $R(Q_{ij}, r_{ij})$ are appearing and on the right-hand side sets of the form $R^c(Q'_{\ell k}, r'_{\ell k})$. So, if (24) cannot be valid for any array (Q_{ij}) and $(Q'_{\ell k})$ (equivalently: (DS, τ) is connected) it means that no Boolean retrieval, consisting of elementary queries exists that is equal to any Boolean query of NOTs of elementary queries. The same argument goes for τ''. In both cases, we use the fact that *any* Boolean query consisting of a finite number of ANDs and ORs can be reduced to the form:

$$\underset{j=1}{\overset{m}{\mathrm{OR}}} \left(\underset{i=1}{\overset{m}{\mathrm{AND}}}\, Q_{ij} \right)$$

(cf. (8)) (or with NOT Q_{ij} in the other case). This was already noted in [3] and mentioned in the Introduction. Its proof is also based on the next lemma. $\blacksquare$

LEMMA. *(See [5, p. 25].) Let $\{B_\alpha \parallel \alpha \in A\}$ be a family of sets and assume that $\{A_\lambda \parallel \lambda \in \mathcal{L}\}$ is a partition of A with each $A_\lambda \neq \phi$. Let $T = \prod_{\lambda \in \mathcal{L}} A_\lambda$. Then*

$$\bigcap_{\lambda \in \ell} \left(\bigcup_{\alpha \in A_\lambda} B_\alpha \right) = - \bigcup_{t \in T} \left(\bigcap_{\lambda \in \ell} B_{t(\lambda)} \right),$$

where $t(\lambda) \in A_\lambda$ and $t = (t(\lambda))_{\lambda \in \ell}$. By taking complements one also has

$$\bigcup_{\lambda \in \ell} \left(\bigcap_{\alpha \in A_\lambda} B_\alpha \right) = \bigcap_{t \in T} \left(\bigcup_{\lambda \in \ell} B_{t(\lambda)} \right).$$

In other, more clear, terms: any $\cap\cup$ of sets A_{ij} can be interpreted as a $\cup\cap$ of the same sets (but in another order) and vise-versa.

REFERENCES

1. D.M. Everett and S.C. Cater, Topology of document retrieval systems, *Journal of the American Society for Information Science* **43** (10), 658–673 (1992).
2. L. Egghe and R. Rousseau, Everett and Cater's retrieval topology, Letter to the editor, *Journal of the American Society for Information Science* **48** (5), 479–480 (1997).
3. L. Egghe and R. Rousseau, Topological aspects of information retrieval, *Journal of the American Society for Information Science* (to appear)(1998).
4. A. Császár, General topology, In *Disquisitiones Mathematicae Hungaricae 9*, Akadémiai Kiadó, Budapest, (1978).
5. J. Dugundji, *Topology*, Allyn and Bacon, Boston, (1966).
6. E. Kreyszig, *Introductory Functional Analysis with Applications*, J. Wiley and Sons, New York, (1978).
7. A. Wilansky, *Topology for Analysis*, Xerox College Publishing, Lexington, MA, (1970).
8. S. Willard, *General Topology*, Addison-Wesley, Reading, MA, (1970).