Everett and Cater's retrieval topology

Peer-reviewed author version

EGGHE, Leo & ROUSSEAU, Ronald (1997) Everett and Cater's retrieval topology.
In: Journal of the American Society for Information Science, 48(5), p. 479-481.

DOI: 10.1002/(SICI)1097-4571(199705)48:5<479::AID-ASI17>3.0.CO;2-U
Handle: http://hdl.handle.net/1942/806

Letter to the editor

Everett and Cater's retrieval topology

Sir,

In this letter we demonstrate that there is an error in Everett and Cater's article about retrieval topologies (Everett & Cater,1992). To make this note self-contained we recall the following definitions adapted from (Everett & Cater,1992).

A retrieval model is a triple (DS,QS,sim), consisting of a document space DS, a query space QS, and a non-negative real-valued *similarity function*

$$\text{sim} : DS \times QS \to \mathbb{R}^+$$

The retrieval topology is the smallest topology on DS which contains all the sets of the form $R(Q,r) = \{D \in DS: \text{sim}(D,Q) > r\}$, with $Q \in QS$ and $r \in \mathbb{R}^+$. Recall that elements of a topology are called open sets. The empty set and the space itself are always open.

Two retrieval models (DS,QS,sim_1) and (DS,QS,sim_2) are said to be essentially equivalent if for every $Q \in QS$: $\text{sim}_1(D_1,Q) < \text{sim}_1(D_2,Q)$ if and only if $\text{sim}_2(D_1,Q) < \text{sim}_2(D_2,Q)$.

Everett and Cater formulate the following Lemma and Theorem:

Lemma 1. For any $Q \in QS$ and $t \in [0,1]$, the set $U(Q,t) = \{D; \text{sim}(D,Q) < t\}$ is open in the retrieval topology.

Theorem 4. Let T_1 and T_2 be the retrieval topologies of two essentially equivalent retrieval models, (DS,QS,sim_1) and (DS,QS,sim_2). Assume further that DS is compact in T_2, then $T_1 = T_2$.

We will show, by giving a counterexample, that these two assertions are wrong. First, however, we recall the definition of a compact set and prove a result on topological spaces that are compact for the retrieval topology.

Definition. A topological space S is said to be compact if every open covering has a finite subcovering. Here, a covering is a set of subsets such that its union is equal to S.

Proposition
If there exists a document D_0 in DS satisfying the following relation:

$$\forall\ Q \in QS, \exists\ c(Q) > 0, \text{ such that } \forall\ D \in DS: c(Q) = \text{sim}(D_0,Q) \leq \text{sim}(D,Q),$$

then DS is compact for the retrieval topology.
Proof. Consider an open covering $(V_i)_i$ of DS. At least one of these V_i must contain D_0. Hence D_0 belongs to a basic open set of the form

$$\bigcap_{j=1}^{k} R(Q_j, r_j)$$

where $R(Q_j,r_j) = \{D \in DS; sim(D,Q_j) > r_j\}$. As, for every $j = 1,...,k$, $sim(D_0,Q_j) = c(Q_j)$, this means that, for every $j = 1,...,k$, $r_j < c(Q_j)$, and consequently, all the $R(Q_j,r_j)$ are equal to DS. Then their intersection is also equal to DS, and hence also the corresponding set V_i. The singleton $\{V_i\} = \{DS\}$ yields a finite open subcovering of the original open covering.

We are now ready to present the counterexamples. Let DS = N (all natural numbers, inclu-ding zero); QS can be any non-empty set. Let sim_1 be defined as follows: $sim_1(D,Q) = 1/5$ for all $Q \in$ QS, except for one special query Q_1. For this special query $sim_1(n,Q_1)$, $n \in N$, is given by the following table:

n	$sim_1(n,Q_1)$
0	1/5
1	1/4
3	3/8
5	7/16
7	15/32
and so on for the odd numbers	
2	1/2
4	3/4
6	7/8
and so on for the even numbers	

Note that the similarity values for the odd numbers converge to 1/2; the similarity values for the even numbers converge to 1.

The open sets for the retrieval topology T_1 are the following: N, ∅ (the empty set), all natural numbers except zero, all natural numbers except the j smallest odd numbers (j=1,2,...), all natural numbers except the odd numbers and the j smallest even numbers (j=2,3,...). Note that the set consisting of all even numbers {2,4,6,...} is NOT an open set for this retrieval topology.

Now, according to the Lemma the set $U(Q_1,1/4) = \{D; sim_1(D,Q_1) < 1/4\}$ should be open. However, $U(Q_1,1/4)$ is the singleton {0} which is not open in the retrieval topology. This example also shows that the function $sim_1(\cdot, Q_1)$ is not continuous for the retrieval topology. Many other examples of U-sets which are not open in the retrieval topology can be given. The proof of this Lemma provided by Everett and Cater basically shows that the U-sets are open in their (pseudo)metric topology, not in the retrieval topology. As the proof of Theorem 4 uses the Lemma this proof is certainly in error. The next example will show that not only the proof but also the theorem itself is wrong.

We consider the same retrieval model as before but - for the second similarity function - make a slight change to the first one. The similarity function sim_2 is everywhere equal to sim_1, except for the value in $(2,Q_1)$. There sim_2 takes the value 5/8. It is now clear that the models (DS=N, QS,

sim_1) and (DS=N, QS, sim_2) are essentially equivalent. Moreover, as $sim_i(0,Q)$ is 1/5 for every Q, i = 1,2, the set DS = N is compact for the retrieval topology (the point zero plays the role of D_0 from the Proposition).. However, the two retrieval topologies do not coincide. For T_2 (the retrieval topology derived from sim_2), the set consisting of all even numbers = $\{n \in N; sim_2(n,Q_1) > 1/2\}$ is clearly open (an element of T_2) . We noted before that this set is not open in T_1. Hence the two topologies are not equal.

We like to end this letter by stating that these errors do not diminish Everett and Cater's contribution to the theory of information retrieval. Indeed, we intend to further study topologies on document spaces and their implications on the retrieval problem.

Leo Egghe
LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium
and UIA, Informatie- en Bibliotheekwetenschap, Universiteitsplein 1, B-2610, Wilrijk, Belgium
and Ronald Rousseau
KHBO, Industrial Sciences and Technology, Zeedijk 101, B-8400 Oostende, Belgium
and LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium

References

Everett,D.M. and Cater,S.C. (1992). Topology of document retrieval systems. *Journal of the American Society for Information Science, 43,* 658-673.