Individual and Collective Approaches for Searcher Satisfaction in IR

Th.P. van der Weide, P. van Bommel

Department of Information Systems, University of Nijmegen
Toernooiveld 1, NL-6525 ED Nijmegen, The Netherlands
tvdw@cs.kun.nl

Technical Report CSI-R9819, 1998

Abstract

The *incremental searcher satisfaction model* for Information Retrieval has been introduced to capture the relevancy of documents under consideration of documents previously presented. In this paper, different approaches for the construction of increment functions are identified, such as the *individual* and the *collective* approach. The requirements posed by these approaches are examined and evaluated with respect to well-known similarity measures used in IR, such as Inclusion, Jaccard's, Dice's, and Cosine coefficient.

1 Introduction

The *incremental searcher satisfaction model* can be used to consider relevancy of documents in the light of previously presented documents ([13]). Incremental satisfaction (also referred to as novelty or surprise) is a major concern for e.g. state-of-the-art Internet search engines aiming at maximal search support (cf. [1], [4]). It can be used to present documents in such a way that the specific searcher can easily determine which (classes of) documents are appropriate for further investigation. Note that some searchers may appreciate similar documents to be presented twice, while others only want to see documents providing sufficient new information (compared to documents previously presented).

In this paper the incremental model is extended. Different approaches for the construction of increment functions are identified. The idea is to define several ways to compare a given document with the set of documents the searcher has already seen. Two approaches are studied in-depth: the *individual* and the *collective* approach. The requirements posed by these approaches are defined within an axiomatic framework. We show that collective increment functions have a strict nature, posing more requirements than individual increment functions. The principles underlying the incremental model are further examined by confronting abovementioned approaches with existing similarity measures, leading to a specialization hierarchy for similarity requirements.

The incremental model can be used in combination with other techniques in this area, such as document ranking techniques (e.g. [12]) and techniques for visualising relevancy (see e.g. [8]). Although we focus on the kernel of IR relevancy treatment and pay little attention to the user interface, we propose the incremental model to be embedded within systems having interaction features especially suited for IR applications (see e.g. [3], [2]). Furthermore, incremental relevancy can be applied in the area of document summarization. For details about incremental summarization see [6], where a linear combination of two similarity functions is used, one for quickly selecting a set of documents, which is more closely investigated by a second, more accurate similarity function.

The organisation of the paper is as follows. In section 2, the incremental searcher satisfaction model is introduced, including a number of axioms for incremental functions. In section 3, the individual

and collective approaches are defined as an instantiation of a generic form for increment functions. This leads to a number of axioms for similarity measures. Section 4 evaluates the requirements discussed in section 3. Specific similarity measures are evaluated with respect to the axioms previously introduced. We use well-known functions such as Inclusion coefficient, Jaccard's coefficient, Dice's coefficient, and Cosine coefficient, as they are found in the literature (see e.g. [11]).

2 The incremental searcher satisfaction model

The Information Retrieval paradigm is about a person (physical or not) having a need for information, and an information collection from which this need is to be satisfied. Elements of the information collection are referred to as documents or information objects.

2.1 Increment functions

The need for information is satisfied by documents. The need for information thus induces a need for information objects. However, after having been presented some (relevant) documents, the yet unsatisfied part of the information need will induce a different need for information objects.

For example, suppose a searcher is interested in information about A, B and C, which has to be satisfied from the following collection:

- x_1 *All about A and something about C*
- x_2 *About B and C*
- x_3 *About B and A*
- x_4 *All about A*

The first document x_1 contains the requested information about A, and some information about C. After inspecting this document, the need for information thus is restricted to information B and C in a limited extend. This is summarized in table 2.1.

document	*title*	*initial need*	*need after* x_1
x_1	All about A and something about C	++	presented
x_2	About B and C	++	++
x_3	About B and A	++	+
x_4	All about A	++	-

Table 1: The residual need for documents

In this paper, we provide a general model to formalize the information need in terms of a need for information objects, and introduce the incremental searcher model as a framework for so-called increment functions. Several instances of increment functions will be discussed.

It is a main objective of an Information Retrieval System to provide an effective disclosure mechanism for a collection $\mathcal{O}$ of information objects. Effective in this context means: yielding relevant documents in return to a searcher query.

In this paper we present the incremental searcher satisfaction model, or *incremental model* for short. In this model, it is assumed that the need for more documents is influenced by what the searcher already has retrieved from the archive. This can be modelled as a function

$$I : \wp(\mathcal{O}) \times \mathcal{O} \mapsto [0, 1]$$

$I(S, x)$ is interpreted as the increment in searcher satisfaction when document x is presented after set S has already been presented to the searcher. The function I is also referred to as the

increment function. A special case occurs when a document is presented without any previously presented documents. In this case, the increment value $I(\varnothing, x)$ is also reffered to as the document need (denoted as $N(x)$). The set S can also be interpreted as previous personal knowledge of the searcher (sometimes also called a user profile).

The incremental model is especially useful for (very) dynamic and distributed archives, such as the World Wide Web. Firstly, as the increment function allows for real-time calculation. This is in contrast with approaches that try to cluster the retrieval result before presenting the clusters to the searcher. Clustering is only possible after all documents have been obtained. Secondly, for distributed archives recall is not useful as a measure for retrieval quality. We rather use a quality measure which is based on total searcher satisfaction, bypassing the need to have global knowledge of the collections involved (see also [13]).

2.2 Basic axioms

The increment function has to satisfy a number of conditions. We first consider two basic axioms. The first axiom states that presenting a document twice does not add anything. The second axiom expresses that the incremental value of a document can not grow after supplying more documents:

IM1	*law of repetition*	$x \in S$	$\Rightarrow$	$I(S, x) = 0$
IM2	*law of growing knowledge*	$S \subseteq T$	$\Rightarrow$	$I(S, x) \geq I(T, x)$

These axioms are tailored to a conventional retrieval environment. The motivation for these axioms is that we assume that presenting documents has a satisfying (non-increasing) effect on the document need. There may, however, be situations that do not support this underlying assumption. These will not be considered in this paper. Note that an alternative approach would be to have IM1':

IM1' *law of repetition* $I(\{x\}, x) = 0.$

IM1' immediately follows from axiom IM1. On the other hand, IM1 can be derived from IM1' combined with IM2:

Proof:

Suppose $x \in S$, then $\{x\} \subseteq S$, and thus from IM2 it follows: $I(S, x) \leq I(\{x\}, x)$. From $I(\{x\}, x) = 0$ we conclude $I(S, x) = 0$.

The following property is an immediate consequence of axiom IM2:

Lemma 2.1 $I(S, x) \leq N(x)$

Proof:

As $\varnothing \subseteq S$, it follows from IM2 that $I(\varnothing, x) \geq I(S, x)$, and thus $N(x) \geq I(S, x)$.

In other words, the maximal satisfaction which can be obtained from a document x is its information need $N(x)$. If x would be presented after the document set S has already been presented, then the incremental searcher satisfaction $I(S, x)$ is at most this maximal satisfaction. As a consequence of this interpretation, the increment function is also referred to as the *residual information need*, i.e., the restant of the information need after being confronted with S.

Next we isolate the effect of presenting a single document.

Lemma 2.2 $I(S \cup \{y\}, x) \leq I(S, y) + I(\{y\}, x)$

Proof:

From IM2 we conclude $I(S \cup \{y\}, x) \leq I(\{y\}, x)$, and thus also $I(S \cup \{y\}, x) \leq I(S, y) + I(\{y\}, x)$.

If presenting the documents from set S does not affect the amount of new information provided by document x after (also) presenting document y, then all information contained in document y is available in the documents from S.

Lemma 2.3 $I(S \cup \{y\}, x) = I(S, y) + I(\{y\}, x) \Rightarrow I(S, y) = 0$

Proof:

Suppose $I(S \cup \{y\}, x) = I(S, y) + I(\{y\}, x)$. From IM1 it follows that $I(S \cup \{y\}, x) \leq I(\{y\}, x)$, and thus $(I(S, y) + I(\{y\}, x) \leq I(\{y\}, x)$, and thus $I(S, y) \leq 0$.

Corollary 2.1

$I(\{x, y\}, z) \leq I(\{x\}, y) + I(\{y\}, z)$
$I(\{x, y\}, z) = I(\{x\}, y) + I(\{y\}, z) \Rightarrow I(\{x\}, y) = 0$

2.3 Effective knowledge

In this section we introduce a third axiom based on effective knowledge. This axiom is expressed in terms of information containment for documents. Information containment is used as a basis for aboutness in the context of matching information objects with queries. In terms of the incremental model, the information containment relation is defined as:

$$x \subseteq_I y \equiv I(\{y\}, x) = 0$$

where $x \subseteq_I y$ is verbalized as: *the information in x is contained within y*, in the context of the information need represented by I. In the sequel, we will omit the index I, and denote information containment as $\subseteq$. The effect on x of presenting y carries over to more complex situations:

Lemma 2.4 $x \subseteq y \iff \forall_S [I(S \cup \{y\}, x) = 0]$

Proof:

From the righthandside of this equation, the lefthandside immediately follows.

Next suppose $x \subseteq y$, or, $I(\{y\}, x) = 0$. Then from IM2 it follows that $I(S \cup \{y\}) \leq I(\{y\}, x)$, and thus $I(S \cup \{y\}, x) = 0$.

If the information in document x is contained within y, then presenting document y eliminates the need for document x:

Lemma 2.5 $x \subseteq y \wedge y \in S \Rightarrow I(S, x) = 0$

Proof:

Suppose $x \subseteq y$, then $I(\{y\}, x) = 0$. Let $y \in S$, then form axiom IM2 we conclude $I(\{y\}, x) \geq I(S, x)$, and thus $I(S, x) = 0$.

Irrelevant documents (i.e. $N(x) = 0$) do not contain any information that is relevant for the searcher. Such documents thus can be seen as empty-information objects. As a consequence, irrelevant documents have special properties:

Lemma 2.6 $N(x) = 0 \Rightarrow x \subseteq y$

From axiom IM1 it directly follows that the relation $\subseteq$ is reflexive. A third requirement to the incremental function is the containment relation to be transitive as this makes the containment relation a partial order on documents. This partial order plays a vital role in the reasoning process within logical models of Information Retrieval (see [10] or [7]). Transitivity is enforced by the following axiom:

$$\text{IM3} \quad \textit{law of effective knowledge} \quad x \subseteq y \Rightarrow I(S,x) \leq I(S,y)$$

So, if the information from document x is contained within y, then document x can not be more surprising than document y. An immediate corrolary of this axiom is that subdocuments can not be more relevant than superdocuments: $x \subseteq y \Rightarrow N(x) \leq N(y)$. The rule IM3 for information containment is sufficient to guarantee the transitivity of the containment relation:

Lemma 2.7 $x \subseteq y \wedge y \subseteq z \Rightarrow x \subseteq z$

Proof:

Suppose $x \subseteq y \wedge y \subseteq z$. From $x \subseteq y$, we conclude from IM3: $I(\{z\},x) \leq I(\{z\},y)$. From the definition of $y \subseteq z$ we conclude $I(\{z\},y) = 0$. As a consequence, $I(\{z\},x) = 0$, or, $x \subseteq z$.

The implication from IM3 may be reversed:

Lemma 2.8 $\forall_S\,[I(S,x) \leq I(S,y)] \Rightarrow x \subseteq y$

Proof:

Suppose $\forall_S\,[I(S,x) \leq I(S,y)]$. By substituting $\{y\}$ for S, we get: $I(\{y\},x) \leq I(\{y\},y) = 0$. In the latter step, axiom IM1 is applied.

2.4 Independent knowledge

In this section we introduce two final axioms based on independent knowledge. These axioms are expressed in terms of the *not about* relation (see e.g. [14] or [5]). For a given retrieval situation, modelled by increment function I, a document y can be considered to be not about document x, denoted as $x \rfloor_I y$, if:

$$x \rfloor_I y \;\equiv\; I(\{y\},x) = N(x)$$

The relation $x \rfloor_I y$ expresses that presenting document y does not influence the need for document x. The index I will be omitted in the rest of this paper. Irrelevant documents have a special place. In this specific retrieval situation, irrelevant documents do not contain any relevant information. Therefore, irrelevant documents do not handle about anything. As a consequence, presenting such a document can not have any effect on the need for any other document:

Lemma 2.9 $N(x) = 0 \Rightarrow x \rfloor y$

Proof:

Suppose $N(x) = 0$, then $I(\{y\},x) \leq N(x)$ implies $I(\{y\},x) = 0$, and thus $x \rfloor y$.

The nature of the not-about relation is layed down in the following axiom:

$$\text{IM4} \quad \textit{law of independent knowledge} \quad x \rfloor y \Rightarrow I(S \cup \{y\},x) = I(S,x)$$

This axiom expresses that the not-about relation is not affected by presenting more documents. If presenting a set S of documents does not have any effect on the need for a document x, then all documents y from S are not about x:

Lemma 2.10 $I(S,x) = N(x) \wedge y \in S \Rightarrow x \rfloor y$

Proof:

Suppose $I(S,x) = N(x)$, and let $y \in S$.

First note that $\{y\} \subseteq S$, and thus (by IM2) we have $I(\{y\},x) \geq I(S,x) = N(x)$.

On the other hand, $\varnothing \subseteq \{y\}$. Applying IM2 once more yields $N(x) = I(\varnothing,x) \geq I(\{y\},x)$.

As a consequence: $I(\{y\},x) = N(x)$, and thus $x \rfloor y$.

For relevant documents x, the relations $x \subseteq y$ and $x \rfloor y$ exclude each other. In other words, if x is not about y, then the infomation of x cannot be contained within y:

Lemma 2.11 If $N(x) > 0$, then $x \rfloor y \Rightarrow x \not\subseteq y$.

Proof:

Suppose x is a relevant document. If $x \rfloor y$ then $I(\{y\},x) = N(x)$. As $N(x) > 0$, we conclude that $x \subseteq y$ does not hold.

Next we consider a final axiom in which the not-about relation is combined with the containment relation. If document x is not about y, and the information of document z is contained within y, then obviously x also not about z:

$$\text{IM5} \quad \textit{law of excluded miracle} \quad x \rfloor y \wedge z \subseteq y \Rightarrow x \rfloor z$$

After having introduced the requirements for increment functions, we will present concrete functions in the next section.

3 Fundamentals of increment functions

In this section we introduce some concrete definitions for increment functions. For this purpose, we also consider similarity functions. We show how the increment function may easily be added to an existing IR situation. In such a case, some document need function N and some measure *Sim* for similarity already have been defined. Furthermore, we assume documents are characterized in terms of a set $\mathcal{D}$ of descriptors by the function $\chi : \mathcal{O} \mapsto \wp(\mathcal{D})$.

We consider a query language $\mathcal{Q}$ as a representation mechanism for the need of a searcher for information. For convenience, we assume that descriptors from $\mathcal{D}$ are used for this purpose. The document need function N, associated with information need $q \in \mathcal{Q}$ then is defined as:

$$N(x) = \mathit{Sim}(q, \chi(x))$$

Using these functions, we first introduce a special class of increment functions, based on the similarity of a document to a set of documents. For given N and *SetSim*, we use increment functions of the following form:

$$I(S,x) = N(x)\,(1 - \mathit{SetSim}(S,x))$$

Thus, $1 - \mathit{SetSim}(S,x)$ can be seen as the amount of new information provided by document x compared to set S. Finally, the outcome is scaled into the interval $[0, N(x)]$ (see lemma 2.1). Increment functions of this form have the following property for information containment:

Lemma 3.1 $x \subseteq y \Rightarrow N(x) = 0 \vee \mathit{SetSim}(\{y\},x) = 1$

For the not-about relation these functions yield a set similarity equal to zero:

Lemma 3.2 $x \rfloor y \Rightarrow N(x) = 0 \vee \mathit{SetSim}(\{y\},x) = 0$

In this section we will discuss different approaches for the computation of *SetSim*. In each case, special requirements for the underlying function *Sim* have to be met. These will be evaluated with respect to well-known similarity functions in section 4.

3.1 The individual approach

In this approach, the similarity between a document and a (non-empty) set of documents is measured as the maximal similarity between the document and any instance of this set:

$$Ind(S,x) = \max\left\{ Sim(\chi(x),\chi(y)) \;\middle|\; y \in S \right\}$$

Furthermore, $Ind(\varnothing, x) = 0$. The expression $Ind(S,x)$ provides the maximal similarity between document x and any of the elements from S of previously presented documents. The resulting increment function is denoted as I_i.

Thus $I_i(S,x)$ gives the fraction of the need $N(x)$ for document x not yet being covered by any previously presented document from S. Consequently, for two documents bringing an equal quantity of new information, the more relevant one is displayed before the less relevant one, as one would expect. Otherwise, the most exotic (and therefore probably highly surprising) documents would be presented before relevant ones.

In the sequel of this section we introduce a number of conditions for simiarity functions, which are sufficient to prove that the resulting increment function satisfies the axioms IM1,.., IM5. The first axiom IM1 is satisfied if equality results in similarity:

S1 $Sim(A,A) = 1$

Proof:

Let $x \in S$, then $Ind(S,x) = 1$ (as $Sim(\chi(x),\chi(x) = 1)$, and thus $I_i(S,x) = 0$.

The validity of axiom IM2 is a direct consequence of the nature of the document similarity function Ind:

Proof:

Let $S \subseteq T$, then $Ind(S,x) \leq Ind(T,x)$, and thus $I_i(S,x) \geq I_i(T,x)$.

For axiom IM3 it is required that

S2 $Sim(A,B) = 1 \Rightarrow Sim(A,X) \geq Sim(B,X)$

S3 $Sim(A,B) = 1 \Rightarrow Sim(X,A) \leq Sim(X,B)$

A consequence of property S3 is:

Lemma 3.3 $Sim(\chi(x),\chi(y)) = 1 \Rightarrow N(x) \leq N(y)$

In words: if the characterization of document x is similar to that of document y, then document x can not be more relevant than document y. Using this property, axiom IM3 can be proven as follows:

Proof:

Suppose $x \subseteq y$, then from lemma 3.1 we conclude $N(x) = 0 \vee Ind(\{y\},x) = 1$. The case $N(x) = 0$ is obvious. So suppose $Sim(\chi(x),\chi(y)) = 1$. Then by subsequent application of **S2**, lemma 3.3, and the definition of I_i, we have:

$$\begin{aligned}
I_i(S,x) &= N(x)\left[1 - \max_{p\in S} Sim(\chi(x),\chi(p))\right] \\
&\leq N(x)\left[1 - \max_{p\in S} Sim(\chi(y),\chi(p))\right] \\
&\leq N(y)\left[1 - \max_{p\in S} Sim(\chi(y),\chi(p))\right] \\
&= I_i(S,y)
\end{aligned}$$

Axiom IM4 does not pose extra requirements on the similarity function:

Proof:

Suppose $x \rfloor y$, then from lemma 3.2 we conclude $N(x) = 0 \vee Ind(\{y\}, x) = 0$ The case $N(x) = 0$ is obvious. So suppose $Sim(\chi(x), \chi(y)) = 0$. Then

$$\begin{aligned} Ind(S \cup \{y\}, x) &= \max_{p \in S \cup \{y\}} Sim(\chi(p), \chi(x)) \\ &= \max(Ind(S, x), Sim(\chi(x), \chi(y)) \\ &= Ind(S, x) \end{aligned}$$

From this the result immediately follows.

Finally we consider increment axiom IM5. The individual increment function satisfies this axiom for similarity functions satisfying **S3**.

Proof:

Suppose $x \rfloor y$ and $z \subseteq y$, then $N(x) = 0 \vee Sim(\chi(x), \chi(y)) = 0$ by application of lemma 3.2. The case $N(x)$ is obvious, so suppose $Sim(\chi(x), \chi(y)) = 0$. Application of lemma 3.1 results in $N(z) \vee Sim(\chi(z), \chi(y)) = 1$. If $N(z) = 0$ then from lemma 2.9 we conclude $x \rfloor z$. So suppose $Sim(\chi(z), \chi(y)) = 1$. As a consequence of **S3** this results in $Sim(\chi(x), \chi(z)) \leq Sim(\chi(x), \chi(y))$. From $Sim(\chi(x), \chi(y)) = 0$ we conclude $Sim(\chi(x), \chi(z)) = 0$, and thus $x \rfloor z$.

3.2 The collective approach

In the collective approach, a new document x is compared to a set S of previously presented documents by comparing the characterization of x with a summary of all presented material from S. The summary $\sigma(S)$ of the set is defined as follows:

$$\sigma(S) = \cup_{y \in S} \chi(y)$$

As a consequence, empty summary is $\sigma(\varnothing) = \varnothing$ and extension of summary is given by $\sigma(S \cup \{x\}) = \sigma(S) \cup \chi(x)$. The similarity between a document x and a set S of documents then is defined as

$$Col(S, x) = Sim(\chi(x), \sigma(S))$$

The expression $Col(S, x)$ provides the degree document x is covered by the total of information provided by the elements from S of previously presented documents. The collective increment function is denoted as I_c.

We need I_c to have the basic property of increment functions $I_c(\varnothing, x) = N(x)$ mentioned in section 2.1. In the collective approach this property holds if similarity with the empty set is impossible:

S4 $Sim(A, \varnothing) = 0$

Next we consider the question under what conditions the axioms IM1 to IM5 are satisfied in the collective approach. The first increment axiom IM1 is satisfied if subsets are similar:

S5 $A \subseteq B \Rightarrow Sim(A, B) = 1$

Proof:

Let $x \in S$, then $\chi(x) \subseteq \sigma(S)$, and as a result of **S5** we have $Sim(\chi(x), \sigma(S)) = 1$. As a consequence $I_c(S, x) = 0$.

The second increment axiom IM2 is satisfied by **S3** and **S5**:

Proof:

Let $S \subseteq T$, then $\sigma(S) \subseteq \sigma(T)$, and as a result of **S5** we get $Sim(\sigma(S), \sigma(T)) = 1$. Using **S3** we now have $Sim(\chi(x), \sigma(S)) \leq Sim(\chi(x), \sigma(T))$. By definition of Col this leads to $Col(S, x) \leq Col(T, x)$, and as a consequence $I_c(S, x) \geq I_c(T, x)$.

Next we consider IM3. In this case the requirements for similarity functions are analogous with the individual approach for increment functions. So, for similarity functions satisfying axioms **S2** and **S3**, the collective increment function satisfies IM3.

Proof:

Suppose $x \subseteq y$, then from lemma 3.1 we conclude $N(x) = 0 \vee Col(\{y\}, x) = 1$. The case $N(x) = 0$ is obvious. So suppose $Col(\{y\}, x) = 1$ and rewrite it to $Sim(\chi(x), \chi(y)) = 1$. Then $I_c(S, x) = N(x)\ [1 - Sim(\chi(x), \sigma(S))] \leq N(y)\ [1 - Sim(\chi(x), \sigma(S))]$ as a consequence of lemma 3.3. Using **S2** we can majorize this by $N(y)\ [1 - Sim(\chi(y), \sigma(S))] = I_c(S, y)$.

Next we consider increment axiom IM4. In the collective approach, this axiom is satisfied if dissimilarity can be extended as follows:

S6 $Sim(A, B) = 0 \Rightarrow Sim(A, S \cup B) = Sim(A, S)$

Proof:

Suppose $x \rfloor y$. Then by definition of the not-about relation we have $I_c(\{y\}, x) = N(x)$. then from lemma 3.2 we conclude $N(x) = 0 \vee Col(\{y\}, x) = 0$. The case $N(x) = 0$ is obvious. So suppose $Col(\{y\}, x) = 0$ and rewrite it to $Sim(\chi(x), \chi(y)) = 0$.

Applying **S6** we now get $Sim(\chi(x), \sigma(S) \cup \chi(y)) = Sim(\chi(x), \sigma(S))$ for any S. Following the definition of Col this results in $Col(S \cup \{y\}, x) = Col(S, x)$ and thus $I_c(S \cup \{y\}, x) = N(x)(1 - Col(S, x))$ from which IM4 is derived.

Finally we consider increment axiom IM5. The collective increment function satisfies this axiom for similarity functions satisfying **S3**. In this case the proof is identical to the proof for individual increment functions.

4 Similarity functions

In this section several instances of increment functions are considered. This is done by choosing specific similarity functions as an instantiation of the generic function Sim used in section 3. Each similarity function is evaluated with respect to the axioms for such a function Sim. We use well-known functions such as Inclusion coefficient, Overlap coefficient, Jaccard's coefficient, Dice's coefficient, and Cosine coefficient, as they are found in the literature (see e.g. [11]). The results of this section are summarized in figure 1.

4.1 Inclusion coefficient

We first consider the Inclusion coefficient for similarity. This coefficient normalizes the amount of overlap $A \cap B$ with the size of A. It is given by:

$$\mathsf{Incl}(A, B) = \frac{|A \cap B|}{|A|}$$

in case $A \neq \varnothing$, while $\mathsf{Incl}(\varnothing, B) = 0$. The axioms for this measure are easily verified, with the exception of axiom **S2**. This axiom is not satisfied, for example, suppose $\mathsf{Incl}(A, B) = 1$, which is equivalent with $A \subseteq B$. Then, by taking X as a superset of $B - A$, we get a counterexample. For the proof of **S3** assume $\mathsf{Incl}(A, B) = 1$, or, equivalently, $A \subseteq B$. Then $|X \cap A| \leq |X \cap B|$, and thus $\mathsf{Incl}(X, A) \leq \mathsf{Incl}(X, B)$.

4.2 Jaccard's coefficient

Next, we consider Jaccard's similarity coefficient. This coefficient normalises intersection $A \cap B$ with the corresponding union:

$$\mathsf{Jacc}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

in case either A or B is nonempty. Furthermore, $\mathsf{Jacc}(\varnothing, \varnothing) = 0$. This coefficient satisfies all axioms, except for axioms **S5** and **S6**.

The proof of axioms **S2** and **S3** directly follows from the observation that $\mathsf{Jacc}(A, B) = 1$ is equivalent with $A = B$. Next we show that axioms **S5** and **S6** are not satisfied. For **S5** let $A \subseteq B$. This leads to $|A| \leq |A \cup B|$ and thus $\mathsf{Jacc}(A, B)$ may be less than 1. For **S6** let $\mathsf{Jacc}(A, B) = 0$. This leads to $A \cap B = \varnothing$ which does not guarantee $\mathsf{Jacc}(A, S \cup B) = \mathsf{Jacc}(A, S)$. In particular **S6** is not satisfied if $|B - S| > 0$.

4.3 Dice's coefficient

Next, we consider Dice's similarity coefficient. This coefficient normalises intersection $A \cap B$ with the sum of its constituents:

$$\mathsf{Dice}(A, B) = \frac{2\,|A \cap B|}{|A| + |B|}$$

in case either A or B is nonempty. Furthermore, $\mathsf{Dice}(\varnothing, \varnothing) = 0$. This coefficient satisfies all axioms, except for axioms **S5** and **S6**.

For **S5** let $A \subseteq B$. This leads to $2\,|A| \leq |A| + |B|$ and thus $\mathsf{Dice}(A, B)$ may be less than 1. The counterexample for **S6** is identical with the Jaccard case.

4.4 Cosine coefficient

Next we consider the Cosine coefficient for similarity. This coefficient normalises the intersection $A \cap B$ with the square root of the corresponding product:

$$\mathsf{Cos}(A, B) = \frac{|A \cap B|}{\sqrt{|A| \times |B|}}$$

in case either A or B is nonempty. Furthermore, $\mathsf{Cos}(\varnothing, \varnothing) = 0$. We will show that the Cosine coefficient behaves the same as the coefficients Jacc and Dice.

For axioms **S2** and **S3** it should be noted that, analogously to the case of Jaccard's and Dice coefficient, $\mathsf{Cos}(A, B) = 1$ is equivalent with $A = B$.

Axioms **S5** and **S6** are not satisfied. For **S5** let $A \subseteq B$, leading to $|A| \leq \sqrt{|A| \times |B|}$. For **S6** the consideration is identical with the Jaccard case.

4.5 Overlap coefficient

Finally, we consider the Overlap coefficient for similarity. This coefficient normalises the intersection $A \cap B$ with the minimum cardinality of its arguments:

$$\mathsf{Ovl}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}$$

in case either A or B is nonempty. Furthermore, $\mathsf{Ovl}(\varnothing, \varnothing) = 0$. We will show that the Overlap coefficient partially behaves the same as the Inclusion coefficients. The difference is that the Overlap coefficient does not satisfy **S3** nor **S6**. The counterexamples are easily found.

4.6 Overview of results

In figure 1 an overview is given. From this overview we conclude that most of the similarity coefficients considered here can be used for increment functions in the individual approach. This is caused by the fact that the individual approach does not pose extreme requirements on similarity functions (only axioms **S1**, **S2**, and **S3** are needed). The Inclusion and the .....

For increment functions in the collective approach only the Inclusion coefficient will suffice, as this approach requires all similarity axioms except **S1** (see section 3.2).

		$\mathsf{Incl}(A,B): \frac{\lvert A\cap B\rvert}{\lvert A\rvert}$	$\mathsf{Jacc}(A,B): \frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$	$\mathsf{Dice}(A,B): \frac{2\lvert A\cap B\rvert}{\lvert A\rvert+\lvert B\rvert}$	$\mathsf{Cos}(A,B): \frac{\lvert A\cap B\rvert}{\sqrt{\lvert A\rvert\times\lvert B\rvert}}$	$\mathsf{Ovl}(A,B): \frac{\lvert A\cap B\rvert}{Min(\lvert A\rvert,\lvert B\rvert)}$
S1:	$Sim(A,A)=1$	$A\cap A=A$	$A\cap A=A\cup A$	$2\lvert A\rvert=\lvert A\rvert+\lvert A\rvert$	$\lvert A\rvert=\sqrt{\lvert A\rvert\times\lvert A\rvert}$	$\lvert A\rvert=\lvert A\rvert$
S2:	$Sim(A,B)=1\Rightarrow$ $Sim(A,X)\geq Sim(B,X)$	No! Let: $B-A\subseteq X$	$A=B$	$A=B$	$A=B$	No! Let: $B\subset A, B\neq A,$ $X\cup A=X\cup B,$ $\lvert X\rvert>\lvert A\rvert$
S3:	$Sim(A,B)=1\Rightarrow$ $Sim(X,A)\leq Sim(X,B)$	$A\subseteq B$	$A=B$	$A=B$	$A=B$	No! Let: $\lvert A\rvert<\lvert X\rvert<\lvert B\rvert$
S4:	$Sim(A,\varnothing)=0$	$\mathsf{Incl}(A,\varnothing)=0$	$\mathsf{Jacc}(A,\varnothing)=0$	$\mathsf{Dice}(A,\varnothing)=0$	$\mathsf{Cos}(A,\varnothing)=0$	$\mathsf{Ovl}(A,\varnothing)=0$
S5:	$A\subseteq B\Rightarrow$ $Sim(A,B)=1$	$\lvert A\cap B\rvert=\lvert A\rvert$	No! $\lvert A\rvert\leq\lvert A\cup B\rvert$	No! $2\lvert A\rvert\leq\lvert A\rvert+\lvert B\rvert$	No! $\lvert A\rvert\leq\lvert\sqrt{A\times B}\rvert$	$A\subseteq B$
S6:	$Sim(A,B)=0\Rightarrow$ $Sim(A,S\cup B)=Sim(A,S)$	$A\cap B=\varnothing$	No! Let: $\lvert B-S\rvert>0$	No! Let: $\lvert B-S\rvert>0$	No! Let: $\lvert B-S\rvert>0$	No! Let: $\lvert S\rvert<\lvert A\rvert<\lvert S\cup B\rvert$

Figure 1: Overview of similarity functions

5 Conclusions

In this paper the incremental searcher satisfaction model for Information Retrieval has been extended. Different approaches for the construction of increment functions were studied. This resulted in a specialization hierarchy for similarity requirements. Following the guidelines for subtyping known from the area of Information Modelling (see e.g. [9]), this specialization hierarchy is graphically shown in figure 2.

The IM-axioms for increment functions resulted in S-axioms for similarity functions. In figure 2 the top category *universal* corresponds with the entire collection of S-axioms. In the right part, the categories *collective* and *individual* correspond with the requirements for collective and individual increment functions, respectively. The intersection of individual and collective requirements, the common subtype *similar*, expresses the specific requirements for cases in which $Sim(A,B)=1$ given by **S2** and **S3**. The category *extended individual* extends the requirements for individual increment functions. This is called a supertype of category *individual*. We have shown that the Cosine, Jaccard's, and Dice's coefficients belong to this category.

In the left part of figure 2 we see a category corresponding with the requirements satisfied by the Inclusion coefficient. We have shown that this category is neither a subtype nor a supertype of the individual and collective requirements. It has two relevant subtypes. The first corresponds with the requirements satisfied by the Overlap coefficient, while the second is the intersection of category *inclusion* with *extended individual*.

In future research attention will be focussed on other approaches for the construction of increment functions. Also, more advanced similarity coefficients will be examined. Besides theoretical models, the incremental searcher satisfaction model will be tested using the incremental satisfaction toolkit which is currently under development.

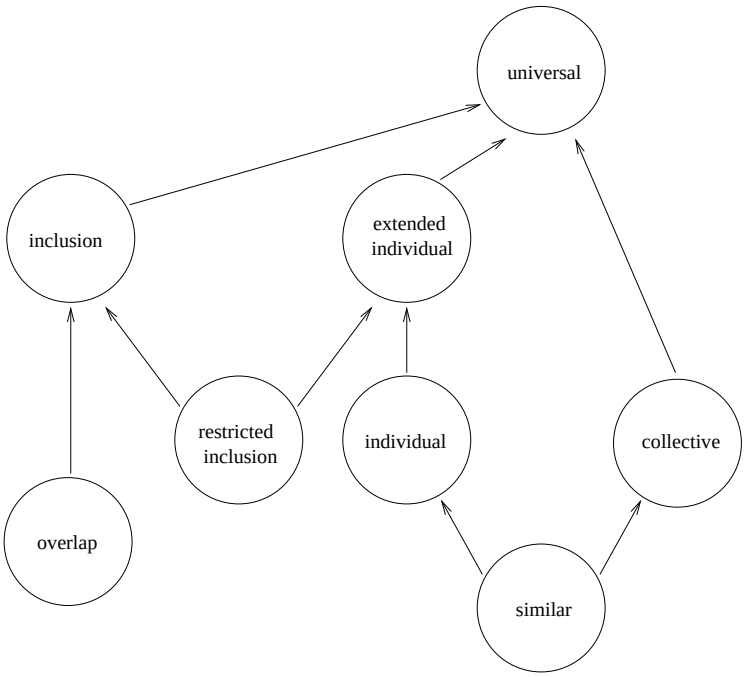

Figure 2: Specialization hierarchy for similarity requirements

References

[1] M. Bates. Where should the person stop and the information search start? *Information, Processing and Management*, 26(5):575–591, 1990.

[2] N. J. Belkin, C. Cool, A. Stein, and U. Thiel. Cases, scripts and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9(3):379–395, 1995.

[3] N.J. Belkin, P.G. Marchetti, and C. Cool. BRAQUE: Design of an interface to support user interaction in information retrieval. *Information Processing & Management*, 29(3):325–344, 1993.

[4] F.C. Berger and P. van Bommel. Personalized Search Support for Networked Document Retrieval Using Link Inference. In R.R. Wagner and H. Thoma, editors, *Proceedings of the 7th International Conference DEXA'96 on Data Base and Expert System Applications*, volume 1134 of *Lecture Notes in Computer Science*, pages 802–811, Zurich, Switzerland, September 1996. Springer-Verlag.

[5] P.D. Bruza and T.W.C. Huibers. Investigating aboutness axioms using information fields. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 112–121, Dublin, July 1994. Springer-Verlag, Berlin.

[6] J. Carbonell, Y. Gang, and J. Goldstein. Automated Query-Relevant Summarization and Diversity-Based Reranking. In I. Ferguson, editor, *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pages 9–14, Nagoya, Japan, August 1997.

[7] Y. Chiarmarella and J.P. Chevallet. About Retrieval Models and Logic. *The Computer Journal*, 35(3):233–242, 1992.

[8] N. Denos, C. Berrut, and M. Mechkour. An image system based on the visualization of system relevance via documents. In *(DEXA 97)*, pages 379–395, 1997.

[9] T.A. Halpin. *Conceptual Schema and Relational Database Design*. Prentice-Hall, Sydney, Australia, 2nd edition, 1995.

[10] S. Krauss, D. Lehmann, and M. Magidor. Nonmonotonic Reasoning, Preferential Models and Cumulative Logics. *Artificial Intelligence*, 44:167–207, 1990.

[11] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, United Kingdom, 1990.

[12] I. Tomek and H. Maurer. Helping the user to select a link. *Hypermedia*, 4(2):111–122, June 1992.

[13] Th. P. van der Weide, T.W.C. Huibers, and P. van Bommel. The Incremental Searcher Satisfaction Model for Information Retrieval. Technical Report CSI-R9719, Computing Science Institute, University of Nijmegen, Nijmegen, The Netherlands, November 1997.

[14] Y.Y. Yao. Measuring Retrieval Effectiveness Based on User Preference of Documents. *Journal of the American Society for Information Science*, 46(2):133–145, 1995.

Contents