Theoretical Benchmarks of evaluation methodologies in XML Retrieval

Tobias Blanke
Queen Mary College, University of London
London, United Kingdom
tobias@dcs.qmul.ac.uk

Mounia Lalmas
Queen Mary College, University of London
London, United Kingdom
mounia@dcs.qmul.ac.uk

ABSTRACT

According to INEX, the evaluation initiative for XML retrieval, the aim of XML retrieval is to retrieve not only relevant document components, but those at the right level of granularity, i.e. document components that specifically answer a query. This paper investigates the use of theoretical benchmarks to describe the INEX 2004 and 2005 evaluation methodology. Theoretical benchmarks concern the formal representation of qualitative properties of information retrieval models. To this end, a Situation Theory framework for the meta-evaluation of XML retrieval is introduced. This model is then used to exemplify theoretical models of user agents and assessment procedures in INEX 2004 and 2005.

1. INTRODUCTION

According to INEX, the evaluation initiative for XML retrieval [5], the aim of XML retrieval is to retrieve not only relevant document components, but those at the right level of granularity, i.e. those that specifically answer a query. To evaluate how effective XML retrieval approaches are, it is necessary to consider whether the 'right' level is correctly identified. For this purpose, two evaluation criteria have been the basis for INEX to consider the structure when evaluating XML retrieval effectiveness. *Topical exhaustivity* reflects the extent to which the information contained in a document component satisfies the information need. *Component specificity* reflects the extent to which a document component focuses on the information need. In order to capture varying degrees of exhaustivity and specificity, INEX has modelled them using graded scales following an investigation by Kekäläinen and Jarvelin [8]. The aim of this paper is to 'meta-evaluate' the use of these scales employing theoretical benchmarks. This paper will therefore consider the changes in the scales used in INEX 2004 and 2005 from a theoretical point of view. It will relate them to models for agent reasoning, as they are expressed in the INEX quantisation functions which map the graded scales onto scalar values.

7th Dutch-Belgian Information Retrieval Workshop, March 28-29,2007, Leuven, Belgium.

By representing the agent reasoning in a formal logical framework we will be able to relate them to exhaustivity and specificity. As shown in [6] rational agents, whether computer systems or human, have the ability to gather information and reason about this gathered information. Afterwards, we analyse the aboutness decisions behind the graded scales for INEX 2004 and 2005. We will demonstrate how to reason about the changes in the graded scales within a theoretical logic-based system. We will build upon earlier work [2], in which the authors show that such a meta-evaluation of XML retrieval is feasible and can deliver some promising initial results.

2. RELATED WORK

As van Rijsbergen [11] has suggested, given the increasing complexity of the retrieval task due to, for example more complex information units, like XML elements, an experimental approach to information retrieval (IR) should be complemented with a theoretical evaluation technique. A theoretical evaluation can be complementary to an experimental evaluation if it helps to clarify the assumptions of retrieval models and if it can identify the characteristics leading to a particular experimental behaviour.

A theoretical evaluation can be done by using a metatheory, as proposed in previous work based on the logical approach to IR [6]. In 1971, Cooper coined the term 'logical relevance' for an objective view on relevance [4]. Van Rijsbergen and others have expressed the logical relevance in terms of the implication $d \rightarrow q$ [11]. Following Huibers' formalism and approach [6], we call such an implication between query and document 'aboutness'. With aboutness, we aim to theoretically capture (1) how INEX sees user agents concluding aboutness, and (2) how system agents present XML elements as answers that are about a query. The former can be modeled with aboutness decisions reflecting user assessments, while the latter can be laid out as the reasoning behind INEX scales for rewarding systems with respect to their ability to deliver exhaustive or specific answers.

In earlier work, Chiaramella [3] presented a theoretical framework to model structured document retrieval. He demonstrated that in order to describe relevance in structured document retrieval, two implications like the one above from van Rijsbergen should be used: $d \rightarrow q$ modelling exhaustivity and $q \rightarrow d$ modelling specificity. Using his logical meta-model, Chiaramella [3] argued that exhaustivity and specificity are not two independent values. He offers a fetch and browse algorithm to deliver the best results for XML retrieval. The most specific answers in structured docu-

ment retrieval are the result of first fetching the exhaustive answers and afterwards browsing through these answers to narrow down the focus. This assumes, that (1) delivering specific answers is the main objective of any structured document retrieval approach and in particular XML retrieval, and that (2) specificity and exhaustivity judgements are based on the same relevant information that is found in the fetch step and specified in the browse step.

Another paper stressing the primary importance of specificity for XML retrieval from a completely different angle than Chiaramella is [9]. In this paper Ogilvie and Lalmas use extensive statistical tests to perform an analysis of the exhaustivity and specificity dimensions used in two rounds of INEX. Their conclusion was that specificity is a sufficient evaluation measure for XML retrieval. We arrive at similar conclusions from a theoretical perspective in section 5.

This paper investigates the use of theoretical benchmarks to describe the INEX 2004 and 2005 evaluation methodology and look at the relationship of exhaustivity and specificity values in it. To this end, a Situation Theory framework for the meta-evaluation of XML retrieval will be presented in the next section as a formal means to express theoretical benchmarks.

3. SITUATION THEORY FRAMEWORK FOR XML RETRIEVAL EVALUATION

Theoretical benchmarks [12] concern the formal representation of qualitative properties of IR models. These properties are described in terms of supported logical axioms and postulates. We use Situation Theory (ST), developed by Barwise and Perry [1], as our logic-based model for XML retrieval. ST offers a logic of information rather than truth assignments and is therefore closer to real-world applications. ST is a mathematical theory of meaning and information with situations as primitives. Situations are partial descriptions of the world and are composed of information items formalised as infons [6]. For IR modelling, queries and documents are modelled as situations, while infons represent a model's information items like keywords or phrases.

Theoretical benchmarks need formalisms, powerful enough to characterize the fundamental properties of retrieval models. If we consider XML elements and queries to be situations, then $S \,\square\!\!\rightsquigarrow\, T$ means that the document component situation S is about the information need expressed in query situation T. Likewise, $S \,\square\!\!\not\rightsquigarrow\, T$ symbolises that S is not about T. 'Dogs' are not 'cats', and are therefore not about each other. With $\odot$, we formalise the fusion of situations. Preclusion, symbolised by $\perp$, expresses that information clashes and leads to anti-aboutness, for which the symbol $\boxtimes\!\!\rightsquigarrow$ is used. E.g., the information 'night' is anti-about the information 'day', as night and day exclude each other. We know that day and night situation are each other's opposite. $\boxtimes\!\!\not\rightsquigarrow$ stands for the negation of anti-aboutness. We will use it mainly to express the minimum aboutness assumption that the information in two situations does not preclude each other. $\equiv$ states that two situations are equivalent, e.g. two document components contain the same information.

Situations are composed of infons, which can have a polarity indicating whether the described objects are supported by the situation or not. In one of the most simple cases for IR, infons represent relationless keywords to model information items. A simple keyword like 'house' is then represented by $\langle\langle house \rangle\rangle$, a short form for an infon with positive polarity. In this paper, we will only use this shortform, but infons can easily be adjusted to more complex information items [6] to reflect structural constraints in XML retrieval. A simple example for an XML situation as a collection of infons reflecting XML structure is: $\{\langle\langle ElementType, Paragraph, i_3; 1\rangle\rangle, \langle\langle Instantiation, flat, i_3; 1\rangle\rangle\}$. ST is particularly useful for the analysis of XML retrieval, as it can represent both structure and content in a unified model [6], as just seen in the example of the ST representation of an XML document component. We do not need two formal ways to describe structure and document content, although this is not exploited in this paper and would like to keep the examples simple.

In this paper we use the outlined ST benchmarking framework to analyse the evaluation process in INEX. In our framework D stands for the document component situation, and Q for the query situation. Following Chiaramella's distinction [3], $D \,\square\!\!\rightsquigarrow\, Q$ models exhaustivity and $Q \,\square\!\!\rightsquigarrow\, D$ models specificity. In [7] Bruza and Huibers present a general ST aboutness criterion. According to them document D is about Q, if D contains at least one infon i such that situation Q is about infon i. This definition, however, leads to problems if we look at graded scales of relevance measures as in INEX 2004 and 2005. The one common infon i could be an infon expressing structure, possibly itself bearing no information useful to a user. In XML retrieval, two XML situations could share the same infons expressing structure, as they share the same document type definition. This does not mean however that they are about each other, as the following example demonstrates: Let us assume that two documents both consist of one paragraph embedded in a section. Then the ST model of both will have the same infons representing this structure. Furthermore, let us assume that the paragraph in the first document is about dogs, whereas the paragraph in the second document is about cats. Therefore they will not be about each other. Aboutness is a relationship of meaning. For INEX, structure in an XML representation only supports meaning but does not create meaning.

We shall use the idea of 'subsituations' instead of simple infons and shall reformulate the above aboutness criterion as a subsituation-based one. A subsituation is a situation S_i that is part of another situation S, where we count the situation as a part of itself, i.e. a situation is a subsituation of itself. S_i is as a situation a combination of infons [6] that has a meaning in itself. For an XML document, this implies XML elements with content and therefore meaning. In the example above, the existence of a paragraph within a section is not a subsituation. Only with the additional information that the paragraph is about dogs or cats can we have meaning and a subsituation. S_i is called a 'strict' subsituation if it is not S. E.g., a situation $\{\langle\langle house \rangle\rangle, \langle\langle garden \rangle\rangle\}$ has as strict meaningful subsituations $\{\langle\langle house \rangle\rangle\}$ and $\{\langle\langle garden \rangle\rangle\}$. $\{\langle\langle house \rangle\rangle, \langle\langle garden \rangle\rangle\}$ is also a subsituation, but not a strict one. By discriminating *strict* subsituations from *non-strict* subsituations, we are able to differentiate aboutness decisions that demand a completely exhaustive or specific match (non-strict subsituation) from those that only expect a partial match (strict subsituation), which is useful to describe user agent reasoning according to INEX, as we shall see later.

Table 1: Quantisations in INEX 2004

Function	$f(e,s)$	User model
$Strict_4$	$f(e,s) = \begin{cases} 1 & \text{if e=3 and s=3} \\ 0 & \text{otherwise} \end{cases}$	UU
Gen_4	$f(e,s) = \begin{cases} 1 & \text{if (e,s) = (3,3)} \\ 0.75 & \text{if (e,s) } \in \{(2,3),(3,2),(3,1)\} \\ 0.5 & \text{if (e,s) } \in \{(1,3),(2,2),(2,1)\} \\ 0.25 & \text{if (e,s) } \in \{(1,2),(1,1)\} \\ 0 & \text{if (e,s) = (0,0)} \end{cases}$	EU
SOG	$f(e,s) = \begin{cases} 1 & \text{if (e,s) = (3,3)} \\ 0.9 & \text{if (e,s) = (2,3)} \\ 0.75 & \text{if (e,s) } \in \{(1,3),(3,2)\} \\ 0.5 & \text{if (e,s) = (2,2)} \\ 0.25 & \text{if (e,s) } \in \{(1,2),(3,1)\} \\ 0.1 & \text{if (e,s) } \in \{(2,1),(1,1)\} \\ 0 & \text{if (e,s) = (0,0)} \end{cases}$	SU
$AnyRel$	$f(e,s) = \begin{cases} 0 & \text{if (e,s) = (0,0)} \\ 1 & \text{otherwise} \end{cases}$	TU

In order to represent the scale of exhaustivity and specificity in Table 3, we divide the document component and query situations into such subsituations, with $D \equiv D_1 \odot ... \odot D_n$ and $Q \equiv Q_1 \odot ... \odot Q_m$. We assume that if D is an exhaustive answer to Q, then it is due to one of the situations D_i that D is composed of. This is how we go beyond the use of a single infon to decide aboutness. Throughout this paper, we will use a notation D_i to describe subsituations that make D an exhaustive answer, while the rest of the subsituations of D are numbered from 1 to n. Q_i describes subsituations that create specific answers.

Thus, we assume exhaustivity and specificity to be properties of a situation. We look at evaluation criteria like exhaustivity and specificity from a topical aboutness point of view. We take them as concrete properties of information objects which are descriptions of the topics in XML elements and query. Based on subsituations and the assumption that exhaustivity and specificity are properties, we can now formulate a new ST aboutness criterion, which is subsituation-based.

DEFINITION 1. *A situation S is about a situation T if and only if T contains a subsituation T_i such that situation S is about situation T_i.*

With this subsituation-based aboutness criterion, we are able to represent agent reasoning according to INEX in the section 4 and the INEX assessment methodology in the section 5 within a single theoretical framework. We speak of rational agent reasoning to include both system and user reasoning.

4. AGENT REPRESENTATIONS IN INEX QUANTISATIONS

In INEX, the scales for exhaustivity and specificity are mapped onto ratio scales. To this end, INEX uses quantisation functions over the two parameters of exhaustivity and specificity: $f(e,s)$. A single relevance scale [0,1] is the result, as presented in Table 1 for INEX 2004 and Table 2 for INEX 2005. The first two columns in both tables are taken from [9]. Please note that a ? stands for elements that are too small to allow an aboutness conclusion. This value of $f(e,s)$ is new to INEX 2005 reflecting the specific problem with XML document components that are too small to bear information.

Quantisations in INEX reflect the importance attached to exhaustivity and specificity. As such they can be used to describe user agent reasoning about results that system agents should return. E.g.: $Strict_4$ as much as $Strict_5$ only credit highly exhaustive and highly specific elements and thus express very demanding user requirements. Within our ST framework, we have the advantage of being able to express a user's need and a system's attempt to satisfy it within the same framework. Both are reasoning processes that follow rules and axioms. This can be considered to be one of the major advantages of a logical meta-evaluation approach. User assessments are as much as system assessments results of reasoning processes. Both are according to [6] reasoning agents with the ability to come to a conclusion whether a document component is about a query or not. In this section we demonstrate the reasoning of user agents, as we are concerned with the representation of the INEX evaluation methodology. In future work, these reasoning patterns can be used to meta-evaluate XML retrieval systems, similarly to the way Huibers [6] has used other agent reasoning models for flat document retrieval systems.

The quantisation of $Strict_4$ as much as its INEX 2005 equivalent $Strict_5$ simulates those user agents only interested in highly exhaustive and highly specific answers. These *unanimous* users will only be satisfied if aboutness systems return the highest exhaustivity and specificity values [6]. In the following formalisations D_j and Q_j denote one of n unique subsituations of an XML situation such as an XML element or a query. D_i marks the subsituation that determines a component to be an exhaustive answer, while Q_i states that the component is a specific answer. The Unanimous User (UU) will only be happy if she can find nothing

Table 2: Quantisations in INEX 2005

Function	$f(e,s)$	User model
$Strict_5$	$f(e,s) = \begin{cases} 1 & \text{if e=2 and s=1} \\ 0 & \text{otherwise} \end{cases}$	UU
$FullySpec$	$f(e,s) = \begin{cases} 1 & \text{if s = 1} \\ 0 & \text{otherwise} \end{cases}$	SDRU
Gen_5	$f(e,s) = \begin{cases} e * s & \text{if e } \in \{1,2\} \\ 0 & \text{otherwise} \end{cases}$	EU
$GenLifted$	$f(e,s) = \begin{cases} (e+1) * s & \text{if e } \in \{1,2\} \\ s & \text{if e = ?} \\ 0 & \text{otherwise} \end{cases}$	EU
$BinExh$	$f(e,s) = \begin{cases} s & \text{if e } \in \{?,1,2\} \\ 0 & \text{otherwise} \end{cases}$	SDRU

else, but these two subsituations D_i and Q_i. She wants them to be equivalent to the situations D and Q, respectively, in order to conclude either $D \boxdot\!\!\rightsquigarrow Q$ or $Q \boxdot\!\!\rightsquigarrow D$.[1]

Unanimous User (UU)

$$\frac{D_i \boxdot\!\!\rightsquigarrow Q, D_i \equiv D, Q_i \boxdot\!\!\rightsquigarrow D, Q_i \equiv Q}{D \boxdot\!\!\rightsquigarrow Q, Q \boxdot\!\!\rightsquigarrow D}$$

A user looking for specific answers but at the same time not wanting to entirely lose out on exhaustivity can be called a Specificity-Oriented User (SU) represented by *SOG* in INEX 2004, but without a real equivalent in INEX 2005. *SOG* only gives *preferences* to specificity by assigning higher quantisation values to higher specificity values.

Specificity-Oriented User (SU)

$$\frac{D_1 \boxtimes\!\!\not\rightsquigarrow Q, ..., D_n \boxtimes\!\!\not\rightsquigarrow Q, Q_i \boxdot\!\!\rightsquigarrow D, Q_i \equiv Q}{D \boxtimes\!\!\not\rightsquigarrow Q, Q \boxdot\!\!\rightsquigarrow D}$$

The complement to *SOG* with a tendency to favouring exhaustivity is Gen_4. It values higher exhaustivity and represents the Exhaustivity-Oriented User (EU). As long as most aspects of the query are discussed, the focus is secondary. The Exhaustivity-Oriented User (EU) does not neglect specificity fully. The focus however is to have $D \boxdot\!\!\rightsquigarrow Q$. For INEX 2005 Gen_5 and $GenLifted$ both place a stronger emphasis on exhaustivity and their ST representation reflects this by demanding $D \boxdot\!\!\rightsquigarrow Q$ as an overall conclusion and rewarding those XML elements that include exhaustivity subsituations.

Exhaustivity-Oriented User (EU)

$$\frac{D_i \boxdot\!\!\rightsquigarrow Q, D_i \equiv D, Q_1 \boxtimes\!\!\not\rightsquigarrow D, ..., Q_n \boxtimes\!\!\not\rightsquigarrow D}{D \boxdot\!\!\rightsquigarrow Q, Q \boxtimes\!\!\not\rightsquigarrow D}$$

In INEX 2004, the *AnyRel* function captures the typical user of mass information systems, happy with any relevant component. There is no equivalent in INEX 2005. The Typical User (TU) would like to see any kind of subsituations, allowing to conclude either exhaustivity or specificity. She is not interested in an overall conclusion of $D \boxdot\!\!\rightsquigarrow Q$ or $Q \boxdot\!\!\rightsquigarrow D$, but in partial conclusions indicating either an exhaustive or a specific answer.

Typical User (TU)

$$\frac{D_1 \boxdot\!\!\rightsquigarrow Q}{D \boxdot\!\!\rightsquigarrow Q}, \cdots, \frac{D_n \boxdot\!\!\rightsquigarrow Q}{D \boxdot\!\!\rightsquigarrow Q}, \frac{Q_1 \boxdot\!\!\rightsquigarrow D}{Q \boxdot\!\!\rightsquigarrow D}, \cdots, \frac{Q_n \boxdot\!\!\rightsquigarrow D}{Q \boxdot\!\!\rightsquigarrow D}$$

Instead of a direct equivalent to *SU*, INEX 2005 comes up with two new user types $BinExh$ and $FullySpec$. Both only look for specificity, as long as exhaustivity is not impossible. $BinExh$ is not as strict with respect to the exhaustivity value. In this sense it corresponds to Chiaramella's [3] earlier suggestions that describe the focus of the answer as the specific interest of XML retrieval. Chiaramella has demonstrated within a theoretical experiment that Structured Document Retrieval Users (SDRU) are interested in specificity as long as the answer remains exhaustive enough. This is why we call this model SDRU:

Structured Document Retrieval User (SDRU)

$$\frac{D_1 \boxtimes\!\!\not\rightsquigarrow Q, ..., D_n \boxtimes\!\!\not\rightsquigarrow Q, Q_i \boxdot\!\!\rightsquigarrow D, Q_i \equiv Q}{Q \boxdot\!\!\rightsquigarrow D}$$

SDRU's differ from SU's only in that their overall conclusion is only influenced by specificity. SDRU's are looking to find a $Q_i \boxdot\!\!\rightsquigarrow D$ in order to conclude $Q \boxdot\!\!\rightsquigarrow D$. Not all users of XML retrieval systems have to be SDRU's, but the particular interest of XML retrieval compared with flat document retrieval is better represented by SDRU's than by other agent models, as the overall conclusion is focused on specificity only.

In this section, we have presented agent reasoning models, as expressed in the INEX quantisations for XML retrieval, based on Chiaramella's differentiation of $D \boxdot\!\!\rightsquigarrow Q$ and $Q \boxdot\!\!\rightsquigarrow D$. We have added a third column to Tables 1 and 2 to summarize these results. We have shown the new focus in INEX 2005 on specificity and would like to investigate this issue further by looking at the transition in terms of the system agents' rewards from INEX 2004 to INEX 2005. The next section places the INEX exhaustivity and specificity assessment scales into the context of ST. We will show that system agents are rewarded if they reflect the

[1] $\frac{S}{T}$ means that if the assumption S is valid, then also the conclusion T.

Table 3: INEX 2004 exhaustivity and specificity situations

Scale	*Exhaustivity* $D \ \square\!\!\rightsquigarrow Q$	*Specificity* $Q \ \square\!\!\rightsquigarrow D$
0	$\frac{D_1 \ \square\!\!\not\rightarrow Q, ..., D_n \ \square\!\!\not\rightarrow Q}{D \ \square\!\!\not\rightarrow Q}$	$\frac{Q_1 \ \square\!\!\not\rightarrow D, ..., Q_n \ \square\!\!\not\rightarrow D}{Q \ \square\!\!\not\rightarrow D}$
1	$\frac{D_1 \boxtimes\!\!\not\rightarrow Q, ..., D_n \boxtimes\!\!\not\rightarrow Q}{D \ \square\!\!\rightsquigarrow Q}$	$\frac{Q_1 \boxtimes\!\!\not\rightarrow D, ..., Q_n \boxtimes\!\!\not\rightarrow D}{Q \ \square\!\!\rightsquigarrow D}$
2	$\frac{D_1 \boxtimes\!\!\not\rightarrow Q, ..., D_n \boxtimes\!\!\not\rightarrow Q, D_i \ \square\!\!\rightsquigarrow Q}{D \ \square\!\!\rightsquigarrow Q}$	$\frac{Q_1 \boxtimes\!\!\not\rightarrow D, ..., Q_n \boxtimes\!\!\not\rightarrow D, Q_i \ \square\!\!\rightsquigarrow D}{Q \ \square\!\!\rightsquigarrow D}$
3	$\frac{D_i \ \square\!\!\rightsquigarrow Q, D_i \equiv D}{D \ \square\!\!\rightsquigarrow Q}$	$\frac{Q_i \ \square\!\!\rightsquigarrow D, Q_i \equiv Q}{Q \ \square\!\!\rightsquigarrow D}$

user agent reasonings. E.g., in order to reach the highest values for exhaustivity and specificity, they must support the reasoning of unanimous users.

5. EXHAUSTIVITY AND SPECIFICITY ASSESSMENTS IN INEX 2004 AND 2005

In this section, we shall first present the reasoning behind INEX 2004 and 2005 evaluation within a ST framework, to afterwards relate the user models and quantisations to this reasoning showing that exhaustivity and specificity are two sides of the same aboutness relation and not two independent criteria. We shall follow a similar argument as in [9], where the authors argue for a focus on specificity as this is the specific interest in XML retrieval and sufficient to evaluate it. INEX 2006 has decided to use only specificity to measure retrieval effectiveness.

We continue our modelling of how an agent perceives how exhaustive and specific a component is. We argue that an agent, either a system or a user, will assess the relevance of a component according to the information contained in both Q and D. In the next two subsections, we will develop two tables representing reasoning models according to INEX 2004 and 2005 definitions of graded scales of relevance, respectively. In the third subsection, we will explore the relationship of exhaustivity and specificity for INEX 2005 and argue for a better integrated assessment showing that, for both specificity and exhaustivity, it is the same relevant information that decides on its values.

5.1 INEX 2004 reasoning

Table 3 describes the INEX 2004 reasoning with a subsituation based aboutness criterion. We build upon earlier work in [2] that has used ST to formally represent assessment decisions in INEX 2004. For scale 0, we cannot find any subsituation that would justify an aboutness conclusion. Scale 1 states that we are undecided whether we can call this aboutness. For this scale, there is no strict subsituation that would allow us to conclude aboutness. For scale 2, D_i is a strict subsituation that makes D exhaustive, while the rest of the subsituations of D are numbered from 1 to n. For 3, there is no other information in the assumption than the subsituation having the property exhaustivity or specificity. Users expect to see only information that is relevant, i.e. $Q_i \equiv Q$ for specificity and $D_i \equiv D$ for exhaustivity.

Table 3 corresponds to the in section 4 defined user agent models if systems are rewarded according to the degree they are able to match user reasoning. An UU agent model will be best represented by a system agent model that combines the assumptions of no other information than relevant one for both exhaustivity and specificity. In order to support a SU, a value of at least 1 for exhaustivity is required to be delivered by the system agent. Within the reasoning of the system agent, the assumptions $D_1 \boxtimes\!\!\not\rightsquigarrow Q, ..., D_n \boxtimes\!\!\rightsquigarrow / \ Q$ must be achieved. Exhaustivity must not have the value 0, as this would make specificity have a value of 0 as well. Analogously, for a system agent to reflect an EU the specificity value counts only as far as it is not 0. The TU is not really represented in the INEX 2004 exhaustivity and specificity situations. Her conclusions are binary and not scaled - either an answer is relevant or not. Therefore the non-representation of the Typical User does not affect our representation and is rather an indication that it is separate from the other INEX agent reasoning models. The Typical User will not appear again in later INEX assessments. For a logical reasoning model also difficult to discriminate are degrees of reasoning. In particular, without extension to our ST framework it seems impossible to discriminate an assessment of 1 or 2. We will later see that this does not pose a problem, as for INEX 2005 assessments the two middle-valued assessments are merged into one.

Let us briefly discuss examples of how such agent reasonings behind the INEX 2004 scales are reflected in combined exhaustivity, specificity assessments. For demonstration purposes, we focus on combinations of very high expectations for specificity or exhaustivity. A combined assessment of (3,3) clearly means an exact match, as $D_i \equiv D$ according to the exhaustivity judgement's assumptions, as well as $Q_i \equiv Q$ according to the specificity judgement's assumptions. Furthermore, (3,2) implies that the subsituation Q_i must be about the subsituation D_i: $Q_i \ \square\!\!\rightsquigarrow D_i$, with $Q_i \ \square\!\!\rightsquigarrow D$ according to the assumptions about a scale 2 specificity judgement and $D_i \equiv D$ according to full exhaustivity. This insight is confirmed by an example, where the query is $\langle\langle house\rangle\rangle$ and $\langle\langle garden\rangle\rangle$, while the document component has $\langle\langle house\rangle\rangle$, $\langle\langle garden\rangle\rangle$, and $\langle\langle car\rangle\rangle$. For this example $D \equiv D_i \odot D_1$, with $D_i \equiv \{\langle\langle house\rangle\rangle, \langle\langle garden\rangle\rangle\}$, $D_1 \equiv \{\langle\langle car\rangle\rangle\}$, and $\odot \equiv \cup$. For the purpose of the example, we assume situation fusion to be a simple union as we assume that the information in the two subsituations is semantically unrelated. Situation composition stands in ST for unrelated information [7], while situation fusion $\odot$ describes related information. Then, the subsituation Q_i must be about D_i and must be $\{\langle\langle house\rangle\rangle, \langle\langle garden\rangle\rangle\}$. A combined assessment of (1,3) implies that none of the other subsituations of the exhaustivity judgement contradicts the information in Q_i, with $D_k \boxtimes\!\!\not\rightarrow Q$ and $Q_i \equiv Q$.

In this subsection we have shown how a subsituation-

Table 4: INEX 2005 exhaustivity and specificity situations

E-Scale	*Exhaustivity* $D \ \Box\rightsquigarrow Q$	*Specificity* $Q \ \Box\rightsquigarrow D$	*S-Scale*
0	$\frac{D_1 \ \Box\not\rightsquigarrow Q, ..., D_n \ \Box\not\rightsquigarrow Q}{D \ \Box\not\rightsquigarrow Q}$	$\frac{Q_1 \ \Box\not\rightsquigarrow D, ..., Q_n \ \Box\not\rightsquigarrow D}{Q \ \Box\not\rightsquigarrow D}$	0
1	$\frac{D_1 \ \boxtimes\not\rightsquigarrow Q, ..., D_n \ \boxtimes\not\rightsquigarrow Q, D_i \ \Box\rightsquigarrow Q}{D \ \Box\rightsquigarrow Q}$	$\frac{Q_1 \ \boxtimes\not\rightsquigarrow D, ..., Q_n \ \boxtimes\not\rightsquigarrow D, Q_i \ \Box\rightsquigarrow D}{Q \ \Box\rightsquigarrow D}$	$\frac{\lvert Q_i\rvert}{\lvert D\rvert}$
2	$\frac{D_i \ \Box\rightsquigarrow Q, D_i \equiv D}{D \ \Box\rightsquigarrow Q}$	$\frac{Q_i \ \Box\rightsquigarrow D, Q_i \equiv Q}{Q \ \Box\rightsquigarrow D}$	1

based aboutness criterion can be used to formalize INEX 2004 assessment decisions. In the next subsection we present the transition from INEX 2004 to INEX 2005 where the focus is much more on specificity.

5.2 INEX 2005 reasoning

According to several studies investigating agreements in the relevance assessments [13], the discrimination of scale 1 and 2 in INEX 2004 does not add value and could potentially lead to confusion. This is confirmed looking at it from a subsituation-based aboutness point of view: For INEX 2004 in Table 3, either a subsituation Q_i or D_i exists (scale 2 and scale 3) or not (scale 0 and scale 1) and if it exists is is either a strict subsituation (scale 2) or the complete situation (scale 3). Three-valued scales cover all the differences in a subsituation-based aboutness, with 0 meaning no subsituation for relevance exists, 1 meaning a strict subsituation exists and 2 meaning the complete XML situation is relevant. INEX 2005 [9] has for exhaustivity a three-valued scale like this.

For specificity, a continuous scale is applied in INEX 2005 with values in [0,1], where 1 represents a fully specific component. The specificity value is derived as follows: in a first phase, assessors highlight text fragments containing relevant information so that in the end, each XML element will have highlighted parts and non-highlighted parts where of course the complete XML element can also be either totally highlighted or not highlighted at all. In a second step the ratio of highlighted text and total text per XML element delivers a specificity value between 0 and 1. This procedure was adopted following the outcome of studies showing it to be a more natural way to assess relevance with more consistent assessments [10]. This new procedure is interesting, as for the first time in INEX, it fulfills the inherent mathematical relationship between query and document component.

The new INEX 2005 specificity measure was developed according to the formalism in [5] describing specificity as the focus of a document component on the topic in the query. This focus is exemplified by using the relationship of the size of those parts of a document component that are about the query to the size of those that are not about the query. In [5], they describe specificity as the relationship of a topic and a component. As for the aboutness relation a topic is in IR defined by the query, we can use the relationship of query and document component as a specificity measure. Furthermore, instead of using concepts as the carriers of meaning as in [5], we use situations. To do so, we first need a measure of the size of a situation's information.

Let $|S|$ represent such a measure of the information size of a situation S. $|.|$ is a counting measure, as Rijsbergen calls it [14]. In INEX 2005, assessors highlight those parts of the document component that are about the query. In INEX 2005 characters are counted to determine the length of highlighted text and its relation to the total size of the element. By using the character ratio of highlighted and non-highlighted text for the specificity judgement, INEX 2005 demands that the specificity value should be a direct reflection of the counting size of the aboutness situation in relation to the document component situation: $spec = \frac{|Q \ \Box\rightsquigarrow D|}{|D|}$. By committing itself at least for specificity to a strict relationship between highlighted and non-highlighted text, INEX 2005 uses the extent to which the query topic is a subset of the component topic as an aboutness criterion for the specificity assessment. Therefore, the highlighting of the assessors forms an aboutness reasoning, where the highlighted parts describe the subsituation Q_i that makes the document component a specific answer. Thus, the size of $|Q \ \Box\rightsquigarrow D|$ is identical to the size of Q_i producing specificity, as Q_i describes exactly those parts of a document component that are making a component relevant to the query. The specificity value would be: $spec = \frac{|Q_i|}{|D|}$. Obviously, the fraction would be 0, if Q_i would not exist or 1 if $Q_i \equiv D$.

Table 4 summarizes the way INEX 2005 assigns values to agent reasoning. The specificity value is directly linked to the difference of $|Q_i|$ and $|D|$. Regarding exhaustivity, we follow the above explained logic of subsituation aboutness. Then, 0 expresses that we cannot find a subsituation to conclude exhaustivity. For scale 1 such a subsituation exists and for scale 2 D_i is the complete situation. As an example for a combined assessment, the new user type in INEX 2005, the SDRU would like to see that $D_i \ \Box\rightsquigarrow Q_i$, with a specificity value of 2 requiring $Q_i \equiv Q$ and an exhaustivity value of at least 1 demanding $D_i \ \Box\rightsquigarrow Q$. $D_i \ \Box\rightsquigarrow Q_i$ is a requirement to satisfy the SDRU, which proves Chiaramella's assumption: In order to achieve the best focus for answers, we have to choose from those XML elements that are exhaustive answers. Their subsituations must be about the specificity subsituation.

Even more than Table 3, Table 4 is derived from the above described agent reasoning models, as here the 3-scaled assessments correlate to the idea of subsituation-based aboutness. The UU is still represented best in system agents that deliver just relevant information and nothing else but non-strict subsituations for specificity and exhaustivity. The EU is still favoured by system agents that avoid the value 0 for specificity, but whose reasoning demands at least $Q_i \ \Box\rightsquigarrow D, Q_1 \ \boxtimes\not\rightsquigarrow D, ..., Q_n \ \boxtimes\not\rightsquigarrow D$. In INEX 2005, it becomes clear that this implies finding any kind of subsituation with the exhaustivity property. Lastly, the SDRU demands from system agents any kind of subsituation with the exhaustivity property allowing her to focus on the conclusion of speci-

ficity. Contrary to INEX 2004, all agent models are covered in Table 4 and no problems occur in discriminating the different degrees.

In the last subsection, we will investigate one of the reasons why INEX 2006 has decided to focus only on specificity in order to evaluate retrieval effectiveness. We show that according to INEX 2005 specificity and exhaustivity values are inseparable and represent two views of the same relationship of query and XML element. This becomes clear as a result of INEX 2005, because specificity follows such a well-defined assessment strategy.

5.3 The relation of exhaustivity and specificity in INEX 2005

As exhaustivity and specificity in INEX express the relationship between one document component and one query, they bear a direct mathematical interpretation. Within our ST framework, the factor by which one document component covers only aspects of one query (specificity), and the factor by which one document component is about all aspects of a query (exhaustivity), can be interpreted as follows: Say Q is a query situation, D is one given component situation, and $|D \sqsubset\!\rightsquigarrow Q|$ stands for the counting size to which D is about Q. According to the formalism in [5], exhaustivity can be defined as the degree to which all aspects of the query are covered by comparing the size of the parts of the aboutness relation that make the document component an exhaustive answer with the size of the query topic: $exh = \frac{|D \sqsubset\!\rightsquigarrow Q|}{|Q|}$. As shown specificity is $spec = \frac{|Q_i|}{|D|}$.

The relevant information in the document component is the same for exhaustivity and specificity. Exhaustivity and specificity only differ in representing whether this relevant information covers all aspects of the topic for exhaustivity or whether for specificity the relevant information does not come with irrelevant information in the same document component. Therefore the highlighting of the INEX 2005 assessors for specificity will also have identified the parts of the document components that determine how exhaustively it answers to the topics in the query. Highlighting is about what is relevant. Specificity and exhaustivity are only two different views on how this relevant information relates to other information - either in the query or in the document component.

As already discussed, in INEX 2005 the exhaustivity value is independently chosen from the specificity one.[2] It is not formally based on highlighting. For exhaustivity, assessors are 'free' to choose a value between 0 and 2, while specificity is determined by calculating the relation of highlighted to non-highlighted text. The deliberation that the relevant information stays the same offers a mathematical relationship of the INEX 2005 specificity value towards exhaustivity. Instead of judging exhaustivity without using the highlighted text, an alternative idea for exhaustivity would be to use the second formula from [5]: $exh = \frac{|D \sqsubset\!\rightsquigarrow Q|}{|Q|}$, as for specificity the complement formula $spec = \frac{|Q_i|}{|D|}$ is used. Looking at it from a subsituation-based aboutness point of view, exh has a clear mathematical interpretation similarly to *spec* by relating the counting size of the exhaustivity subsituation to the counting size of the query: $exh = \frac{|D_i|}{|Q|}$. This relation measures the degree to which a document component covers the concepts requested by a topic.

The question of the relation of exhaustivity and specificity values for INEX has been intensely discussed. Ogilvie and Lalmas in [9] go as far as to suggest that after INEX 2005 only specificity should be used for the evaluation, as the specificity-oriented quantisation $BinExh$ can be used to predict exhaustivity-oriented quantisations such as $GenLifted$. This indicates that exhaustivity and specificity are not independent values. Chiaramella [3] even declares that a specificity judgement should be seen as a more narrow focus of an exhaustivity assessment. In his theoretical framework, $Q \sqsubset\!\rightsquigarrow D$ is only evaluated against those document component situations D that are about the query Q: $D \sqsubset\!\rightsquigarrow Q$. We cannot go as far as Chiaramella. By validating our assumption that the same relevant information is used for exhaustivity and specificity assessment within an example, we can however state that these two assessment values are not without relation, but offer two views on the same aboutness relation between query and document component.

In the following example, using $exh = \frac{|D_i|}{|Q|}$ and $spec = \frac{|Q \sqsubset\!\rightsquigarrow D|}{|D|}$ as evaluation measures, we demonstrate that the results are the expected preferences in terms of exhaustivity and specificity. To keep the aboutness relation simple, we assume that the information in query and document components is constituted by their keywords and by their keywords only. Furthermore, we assume that the assessors only use the overlap between these keywords as an aboutness decision. A document component stating 'Dogs are not cats' is relevant to a query about the topic 'cats', as the keyword cats is part of the document component. This keyword will be the only highlighted part in the component. In a ST framework the document component is $\{\langle\langle dogs\rangle\rangle, \langle\langle cats\rangle\rangle\}$ and the query is $\{\langle\langle cats\rangle\rangle\}$. The highlighted text will have 4 characters, the non-highlighted will have 13 characters. In order to simplify our calculations and keep the overview in our example, we assume that the document components only consist of keywords. Then the counting size of a document situation in the INEX 2005 assessment will be the number of characters each keyword information item has plus one whitespace separating the keywords in the text. Like this we avoid confusion about the counting size when describing the document components and query in our example as situations.

Let us therefore assume the following assessment situation: $\{\langle\langle house\rangle\rangle, \langle\langle garden\rangle\rangle, \langle\langle flat\rangle\rangle\}$ is document component situation $D1$ with a counting size of 15. Let $D2$ be $\{\langle\langle house\rangle\rangle, \langle\langle close\rangle\rangle, \langle\langle garage\rangle\rangle\}$ with $|D2| = 16$. We have 5 query situations: $Q1$ is $\{\langle\langle house\rangle\rangle, \langle\langle garden\rangle\rangle\}$, $Q2$ $\{\langle\langle house\rangle\rangle\}$, $Q3$ $\{\langle\langle garden\rangle\rangle\}$, $Q4$ $\{\langle\langle house\rangle\rangle, \langle\langle garden\rangle\rangle, \langle\langle flat\rangle\rangle\}$, and $Q5$ $\{\langle\langle house\rangle\rangle, \langle\langle garden\rangle\rangle, \langle\langle flat\rangle\rangle, \langle\langle car\rangle\rangle\}$. Then $|Q1| = 11$, $|Q2| = 5$, $|Q3| = 5$, $|Q4| = 15$, and $|Q5| = 19$. This example is paradigmatic, as we cover all 4 possible combinations: either the information in the query is fully covered in the document component, or the document component has no other, but not all information of the query, or both document component and query share information, but both also have other information, or query and document component do not share information at all.

Table 5 summarizes the assessment outcomes for $exh = \frac{|D_i|}{|Q|}$ and $spec = \frac{|Q_i|}{|D|}$ in this example. It clearly presents the expected preferences in terms of exhaustivity and specificity

[2]The exception is non-aboutness: A value of 0 for specificity has to mean a 0 for exhaustivity and vice versa.

Table 5: Example with new exhaustivity and specificity measures

	Q1		Q2		Q3		Q4		Q5	
	spec	exh	spec	exh	spec	exh	spec	exh	spec	exh
D1	0.73	1	0.33	1	0.4	1	1	1	1	0.79
D2	0.31	0.45	0.31	1	0	0	0.31	0.33	0.31	0.33

assessments. Also, a 0 assessment in one of the measures leads to a 0 assessment in the other. The example of $Q4$ and $D1$ offers a complete match. The assessment results are always between 0 and 1, as the size of the information overlap of query and document can never be larger than either the counting size of the query situation or the counting size of the component situation. An empty query or an empty document component will lead to an undefined assessment result. For XML retrieval, an empty document component will be an XML element without content, which we defined above as meaningless and therefore no subsituation. We could at this point easily introduce a threshold for the counting size of the subsituation that would exclude those that are too small and therefore meaningless, as it has been done for INEX 2005. Finally, Table 5 shows that it is possible to use the same relevant information as a basis for the exhaustivity and specificity assessments without changing their outcomes in the INEX 2005 assessment procedure. This supports Chiaramella's idea of specificity and exhaustivity as two dependent values.

As a result of this subsection, we do not suggest that $exh = \frac{|D_i|}{|Q|}$ is necessarily a better measure for exhaustivity than the INEX 2005 scale. This could only be done after extensive statistical tests like in [9], which have led to the use of specificity only as an evaluation measure in INEX 2006. By using the same highlighting for exhaustivity that was used for specificity, we theoretically demonstrated that it is possible to look at exhaustivity and specificity as two views of the same property of an aboutness relation and not as two different aboutness relations.

6. CONCLUSION AND FUTURE WORK

This paper presented the potential of a ST based theoretical approach for a meta-evaluation of XML retrieval evaluation standards. We introduced a subsituation-based aboutness criterion that led to an integrated model for user reasoning and assessments methodologies in INEX 2004 and 2005. We could represent how different INEX quantisations express agent models. ST was used to formalise these as reasoning processes. In a second step, we were able to relate these agent models to the INEX evaluation scales and show which patterns of reasoning are involved. We could indicate a theoretically consistent alternative treatment of exhaustivity and specificity for INEX 2005 and suggested to consider both not as independent values.

In future work, we would like to employ these agent models within a ST-based theoretical meta-evaluation of XML retrieval models. Thereby we aim to achieve an integrated framework to represent user and system reasoning.

7. REFERENCES

[1] J. Barwise and J. Perry. *Situations and Attitudes*. MIT Press, Cambridge, MA, 1982.

[2] T. Blanke and M. Lalmas. Theoretical benchmarks of xml retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 613–614, New York, NY, USA, 2006. ACM Press.

[3] Y. Chiaramella. Information retrieval and structured documents. In *Lectures on information retrieval*, pages 286–309. Springer-Verlag, New York, 2001.

[4] W. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7:19–37, 1971.

[5] N. Gövert, G. Kazai, N. Fuhr, and M. Lalmas. Evaluating the effectiveness of content-oriented XML retrieval. *Journal of Information Retrieval*, 9(6):699–722, 2006.

[6] T. W. Huibers. *An Axiomatic Theory for Information Retrieval*. Universiteit Utrecht, Utrecht, 1996.

[7] T. W. C. Huibers and P. D. Bruza. Situations: A general framework for studying Information Retrieval. In R. Leon, editor, *Information retrieval: New systems and current research, Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialists Group*, pages 3–25. Taylor Graham, Drymen, Scotland, 1996.

[8] J. Kekäläinen and K. Järvelin. Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.

[9] P. Ogilvie and M. Lalmas. Investigating the exhaustivity dimension in content-oriented xml element retrieval evaluation. In *15th ACM Conference on Information and Knowledge Management (CIKM 2006)*, Arlington, VA, 2006.

[10] J. Pehcevski and J. A. Thom. Hixeval: Highlighting xml retrieval evaluation. In *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005). Dagstuhl 28-30 November 2005*, pages 43–57. Springer-Verlag, New York, January 2006.

[11] C. J. v. Rijsbergen. Towards an information logic. ACM Press, 1989.

[12] D. Song, K.-F. Wong, P. Bruza, and C.-H. Cheng. Application of aboutness to functional benchmarking in information retrieval. *ACM Trans. Inf. Syst.*, 19(4):337–370, 2001.

[13] A. Trotman. Wanted: Element retrieval users. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 58–64, Glasgow, 2005.

[14] C. J. v. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, 2004.