Working Group Report: the Assessment Tool

Benjamin Piwowarski
LIP 6, Paris, France
bpiwowar@poleia.lip6.fr

ABSTRACT
This paper is the report of the working group on the evaluation assessment interface that was used in INEX'03. This paper describes the changes that are planned for INEX'04 and the different issues that were raised during the working group session.

1. INTRODUCTION
A description of the INEX'03 interface can be found in [1].

This year, the assessment tool was completely redesigned. The first change was the user interface: a single document view was used both to read the document and to assess its components. This change was appreciated by almost every participant. Some enhancements have been suggested (section 2) to ease the assessment process through more user assistance.

The changes were not only cosmetic, as rules ensuring *consistency* and *exhaustivity* of assessments were a main component of the interface. The consistency check (section 3.1) ensures that assessments within the same document are consistent with respect to the definition of the INEX scale. For example, a non relevant element cannot contain relevant elements. The exhaustivity check (section 3.2) ensures that most (if not all) of the highly specific elements are found within assessed documents. Finding highly specific elements is an important point since finding those elements is the goal of an XML information retrieval system. Obtaining consistent and exhaustive relevance assessment is thus crucial for the appropriate comparison of retrieval approaches.

Notations
In this report, an assessment value in the INEX'03 scale is denoted by ExSy (exhaustivity is x, specificity is y), Ex (exhaustivity is x, specificity is unknown) or Sy (specificity is y, exhaustivity is unknown).

2. ENHANCEMENTS
In this section, enhancements that were proposed for the next INEX campaign are described. Every point will be considered when the current interface is extended, but time constraints will possibly postpone some enhancements.

Efficiency
After each assessment, the server (which is actually in Paris) is contacted in order to check the different constraints; its answer updates the document view. This solution was chosen as it was the easiest, but for assessors from distant countries – like e.g. USA, Australia, New Zealand – there was a noticeable delay. Two solutions to this problem are possible:

1. Set up local mirrors;
2. Perform the constraint check on the host (e.g. with javascript) and send the assessments for validation only when leaving the document view.

The first solution is the easiest as it does not involve new development. The second is the best because it allows to centralise all the assessments, but it involves new developments.

Interface
Some participants proposed interface enhancements that would help to speed up or ease the assessment process:

rules When assessing sets of elements, the interface sometimes fail to predict the set of values that those elements can take together. This clearly should not happen.

tree-view An XML tree view of the current document could give a quicker access to distant parts of the structure.

bookmarks When assessing a document, it is often useful to go and look around the element to assess and then come back to this element: bookmarks should allow to do this quicker.

keyword highlighting New highlighting modes like e.g. background, border, font colour in order to distinguish more easily different group of keywords.

New icon set
G. Kazai proposed a new icon set (figure 2) that is more closely related to the INEX'03 scale. Hopefully, the scale will not change next year so we can use them. An empty disc is used to symbolise the "irrelevant" part of the component; a plain disc (shades of blue, from highly to marginally exhaustive) symbolises the "relevant" part of the component.

Exhaustivity \ Specificity	0	1	2	3
0	○			
1				
2				
3				

Figure 1: The new icon set for INEX'04

3. CONSISTENCY AND EXHAUSTIVITY

In this section, consistency and exhaustivity rules are described. In each subsection, rules used for INEX'03 are first exposed. To ensure even more consistency and exhaustivity[1] in INEX'04 assessments, new rules are then proposed. Some of the latter are still to be debated.

In the following, an element is one XML tag while its children includes XML tags *and XML text nodes*. For example, a paragraph with some text within a `<it>` tag will have three children: a text node (before the `<it>`), the `<it>` node and then another text node (after the `</it>`). Even if text nodes cannot be assessed (this is an open issue), they are taken into account while applying the consistency and exhaustivity rules.

3.1 Consistency

The consistency rules ensure a set of assessments within the same document are consistent with respect to the definition of exhaustivity and specificity. They are both used to check an assessment is valid and to infer automatically some assessments. In INEX'03, 7 % of assessments were automatic. An element is automatically assessed when the rules reduces the set of possible assessments to one element: defining new rules not only ensures assessments are more consistent, it is also useful to speed up the assessment process. An element can also be inconsistent when this set is empty. This occurs when some rules change or are added, or when the interface fails to predict the possible choices. The latter can happen when one is assessing a set of elements.

INEX'03

1. The exhaustivity of an element is always superior or equal to the maximum of children exhaustivity. This rule ensure no more relevant information is found in an element than within each of its children.

2. The specificity of an element is inferior or equal to the specificity of any of its child. That rule states that the ratio of relevant information in the element cannot be superior to the ratio of relevant information in its children. For instance, we cannot assess the element S3 if all its children are S2.

New rules

The following rules were not added in INEX'03 due to time constraints, but can be somehow derived from the definition of exhaustivity and specificity, except the third one.

1. The first is the symmetric case of the INEX'03 rule 1. It states that there cannot be more relevant information in an element than in its children: the exhaustivity of an element is inferior or equal to the sum of its children exhaustivity.

2. The ratio of relevant information in an element cannot be inferior to the ratio of relevant information in all its children: the element specificity is superior or equal to the minimum specificity of its children.

3. The last rule is (and was!) heavily discussed. Its role its to ensure that a highly specific element does not have any descendant with the same exhaustivity since it would imply that one of its descendants is as good as the element for an XML information retrieval system to retrieve. This rule is also an extension of the rule 1 in INEX'03: when the element is S3, the exhaustivity is always superior (and not anymore equal) to the maximum of children exhaustivity. The main critic of this rule is that the exhaustivity scale has only three values: the maximum number of elements between the root of the document and any leaf in the XML tree which can be highly specific is thus 3. Furthermore, descendants of an E1S3 element are not relevant with this rule. It should be debated whether this is a too restrictive hypothesis. Another solution would be to restrict the application of this rule to elements assessed E2S3 or E3S3 (and not anymore to elements assessed E1S3).

3.2 Exhaustivity

Exhaustivity rules were much more discussed than consistency rules. The main reason is that consistency rules are somehow *implied* by the definition of exhaustivity and specificity, while exhaustivity is not yet fully understood. The second one is that exhaustivity rules are applied after each assessment and add new elements in the set of assessments to be done. Adding too many elements increase the task burden while adding too few elements does not ensure anymore that we find all S3 elements. The balance between those two extrema is difficult to find.

But the importance of those rules is fully illustrated by this statistic: in INEX'03, 68 % of the S3 elements were not initially in the pools – which implies that adding elements *is* necessary to ensure the exhaustivity of the test collection.

INEX'03

1. When the element is not relevant, nothing is added. This rule is useful since we do want non relevant documents to be assessed as fast as possible – as assessor should concentrate on documents that contains relevant parts.

2. When the element is S3, do not add children but do add ancestors: when a highly specific element is found, there is no need to assess its descendants as this is the only kind of elements we are searching for. This is especially true if we consider the third new consistency rule.

3. Otherwise, add all the children and all the ancestors of the assessed element. This rule is applied when the

[1] exhaustivity is not related to the one of the INEX scale dimension, but to the extent with which all the S3 elements are found

element is neither not relevant, neither highly specific: there is some more specific elements within it that have to be found.

New rules

Only one new rule is planned in order to reduce the number of elements to be added. This rule was obviously one of the most discussed one. The main idea is to prevent any "loss" of relevance between an element and its children, that is to only add the children of a marginally or fairly specific assessed element when there is no children that contain as much relevant information as the assessed element. More precisely, when the sum of the children exhaustivity is superior or equal to the element exhaustivity, no children are added. For example, if an element is assessed E3S2 and that all the relevance of the element is found in one child (that is, one child is E3), there is no need to ask for the assessors to find other relevant parts within the other children of the assessed element – though he can always assess them. Other children are thus removed from the list of elements that have to be assessed.

4. CONCLUSION

Some points that were discussed during this working group were fully debated; this proves the assessment tool is not only a graphical interface, it is also closely related to (1) the assessor effort (2) the quality of the INEX collection (3) the definition of what is relevance. This led to the "I don't wish to assess that" problem which is related to points (1), (2) and (3). What if I really don't want to assess an element? This debate, if I recall well, ended up (or almost) in the definition of a possible new value in INEX scale, namely the "not meaningful" value – the element cannot be judged by itself as it is too small (which implies descendants are also not meaningful?).

The new interface used in INEX'03 will be extended next year to include some of the changes described in this report. Some issues, especially those related to exhaustivity, were much debated in the working group and there is no full agreement upon participants. The new rules will thus be discussed in a forum which is available on the web (http://inex.lip6.fr), along with the possible proposition of new ones.

Eventually, I would like to thank every participant of this working group, feedback being an important part of the development of a good interface for assessments.

5. REFERENCES

[1] G. Kazai, M. Lalmas, and B. Piwowarski. Relevance assessment guide. In *Proceedings of the Second Annual Workshop of the Initiative for the Evaluation of XML retrieval (INEX)*, DELOS workshop, Dagstuhl, Germany, Dec. 2003. ERCIM.