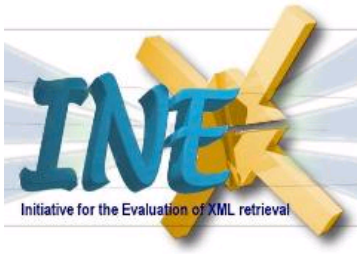


INEX'03 Relevance Assessment Guide

1. Introduction

During the retrieval runs, participating organisations evaluated the 66 INEX'03 topics (36 content-only and 30 content-and-structure queries) against the IEEE Computer Society document collection and produced a list (or set) of document components (XML elements[1]) as their retrieval results for each topic. The top 1500 components in a topic's retrieval results were then submitted to INEX. The submissions received from the different participating groups have now been pooled and redistributed to the participating groups (to the topic authors whenever possible) for relevance assessment. Note that the assessment of a given topic should not be regarded as a group task, but should be provided by one person only (e.g. by the topic author or the assigned assessor).

The aim of this guide is to outline the process of providing relevance assessments for the INEX'03 test collection. This requires first a definition of relevance for XML retrieval (Section 2), followed by details of what (Sections 3) and how (Section 4) to assess. Finally, we describe the on-line relevance assessment system that should be used to record your assessments (Section 5).

2. Relevance dimensions: exhaustivity and specificity

Relevance in INEX is defined according to the following two dimensions:

- ***Exhaustivity (e-value for short),*** *which describes the extent to which the document component discusses the topic of request.*
- ***Specificity (s-value for short),*** *which describes the extent to which the document component focuses on the topic of request.*

To assess exhaustivity, we adopt the following 4-point scale:

> **0: Not exhaustive,** the document component does not discuss the topic of request at all.
> **1: Marginally exhaustive**, the document component discusses only few aspects of the topic of request.
> **2: Fairly exhaustive**, the document component discusses many aspects of the topic of request.
> **3: Highly exhaustive**, the document component discusses most or all aspects of the topic of request.

To assess specificity, we adopt the following 4-point scale:

> **0: Not specific**, the topic of request is not a theme of the document component.
> **1: Marginally specific**, the topic of request is a minor theme of the document component
> **2: Fairly specific**, the topic of request is a major theme of the document component.
> **3: Highly specific**, the topic of request is the only theme of the document component.

A document component can be assessed as highly exhaustive (e-value 3) even if it is not specific to the topic of request – that is, the topic of request can be a major theme (s-value 2) or a minor theme (s-value 1) of the component – as long as all or most aspects of the topic is discussed (e.g. a component may be highly exhaustive to the topic regardless of how much additional, irrelevant information it contains). Similarly, a document component can be assessed as highly specific (s-value 3) even if it discusses many (e-value 2) or only a few (e-value 1) aspects of the topic - as long as the topic of request is the only theme of the component. However, a document component that does not discuss the topic of request at all (e-value 0) must have an s-value of 0, and vice versa.

[1] The terms document component and XML element are used interchangeably.

3. What to judge

Depending on the topic, a pooled result set may contain initially between 500 and 1,500 document components of 500 - 510 articles, where a component may be a title, paragraph, section, or whole article etc.

Traditionally, in evaluation initiatives for information retrieval, like TREC, relevance is judged on document level, which is treated as the atomic unit of retrieval. In XML retrieval, the retrieval results may contain document components of varying granularity, e.g. paragraphs, subsections, sections, articles etc. Therefore, to provide comprehensive relevance assessment for an XML test collection, *it is necessary to obtain assessment for the different levels of granularity*.

This means that if you find, say, a section of an article relevant to the topic of the request, you will then need to provide assessment - both with regards to exhaustivity and specificity - for the found relevant component, for all its ascendant elements until you reach the article component, and for all its descendant elements until you have identified all relevant sub-components.

Such comprehensive assessments are necessary as it is demonstrated by the following example. Consider the XML structure in Figure 1. Let us say that you judged Section C, the document component that encapsulates all text fragments relevant to the topic, as highly exhaustive (e-value 3) and fairly specific (s-value 2). Given only this single assessment it would not be possible to deduce the exhaustivity and specificity levels of the ascending or descending elements. For example, Body D and Article E may be judged fairly or marginally specific depending on the volume of additional, irrelevant information contained within the sections other than Section C. Looking at the sub-components of Section C, it is clear that no conclusions can be drawn from Section C's assessment regarding the exhaustivity or specificity levels of its sub-components. For instance, both Sub-Sections A and B may be marginally, fairly or highly exhaustive, and smaller components, such as Paragraph 3, could even be irrelevant.

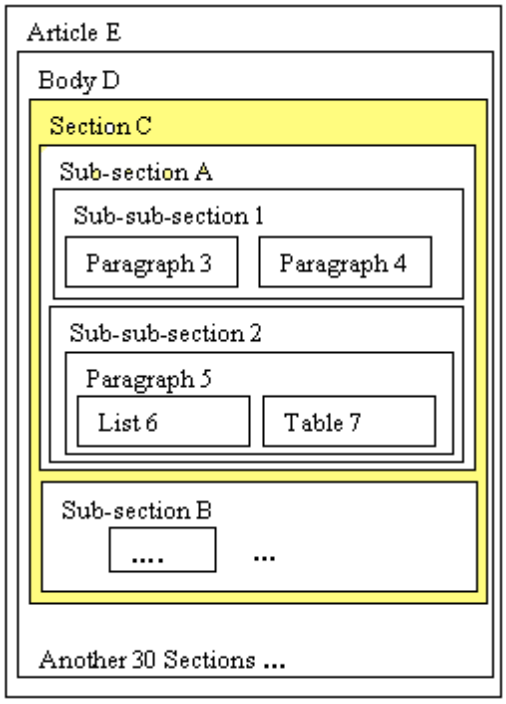


Figure 1. Example XML structure and result element

As a general rule it can be said that the exhaustivity level of a parent element is always equal to or greater than the exhaustivity level of its children elements. This is due to the cumulative characteristics of exhaustiveness. For example, the parent of a highly exhaustive element will always be highly exhaustive since the child element already discusses all or most aspects of the topic. Another rule for the exhaustivity dimension is that the parent of non-exhaustive child elements (i.e. all with e-value 0) will also be not exhaustive (e-value 0). A rule regarding specificity is that an element has an s-value

that is greater than 0 if one of its child elements has an s-value different from 0, and less or equal to the maximum s-value of all its child elements. For instance, suppose that a parent element has tiny child element with s-value 1 and a large child element with s-value 2, then the s-value of that parent element will be 1 or 2. However, besides these general rules, no specific rules exist that would automate all the assessment of ascendant and descendant elements of relevant components. Therefore, you will need to explicitly judge all elements that contain relevant information. This is the only way to ensure both exhaustive and consistent relevance assessments.

4. How to judge

To assess the exhaustivity and specificity of document components, we recommend a three-pass approach.

- During the first pass, you should skim-read the whole article (that a result element is a part of - even if the result element itself is not relevant!) and identify any relevant information as you go along. The on-line system will assist you in this task by highlighting keywords within the article (see Section 5).
- In the second pass, you should assess the exhaustivity and specificity of the relevant components (i.e. identified in the first phase), and that of their ascendant and descendant XML elements.
- To ensure exhaustive assessments, in the third phase, you should assess the exhaustivity and specificity of the descendant XML elements of all elements that have been assessed as relevant during the second phase.

The on-line assessment system (see Section 5) will identify for you all elements that have to be assessed for phases 2 and 3.

During the relevance assessment of a given topic, all parts of the topic specification should be consulted in the following order of priority: narrative, topic description, topic title and keywords. The narrative should be treated as the most authoritative description of the user's information need, and hence it serves as the main point of reference against which relevance should be assessed. In case there is conflicting information between the narrative and other parts of a topic, the information contained in the narrative is decisive. The keywords should be used strictly as a source of possibly relevant cue words and hence only as a means of aiding your assessment. You should not rely only on the presence or absence of these keywords in document components to judge their relevance. It may be that a component contains some or maybe all the keywords, but is irrelevant to the topic of the request. Also, there may be components that contain none of the keywords yet are relevant to the topic. The same applies to the terms listed within the topic title!

In the case of content-and-structure (CAS) topics, the topic titles contain structural constraints in the form of XPath expressions. Although the structural conditions are there to impose a constraint on the structure, you are asked as an assessor to assess the elements returned for a CAS topic as whether they satisfy your information need (as specified by the topic) mainly with respect to the content criterion. Therefore, you should not assess an element as “not relevant” because the structural condition is not satisfied. In fact, your assessment of CAS topic should be very similar to that of content-only (CO) topics, although in the former the structural conditions may influence your assessment (to a small extent).

Note that some result elements are related to each other (ascendant/descendant), e.g. an article and some sections or paragraphs within the article. This should not influence your assessment. For example if the pooled result contains Chapter 1 and then Section 1.3, you should not assume that Section 1.3 is more relevant than Sections 1.1, 1.2, and 1.4, or that Chapter 1 is more relevant than Section 1.3 or vice versa. Remember that the pooled results are the product of different retrieval engines, which warrants no assumptions about the level of relevance based on the number of retrieved related components!

You should judge each document component on its own merits! That is, a document component is still relevant even if it the twentieth you have seen with the same information! It is imperative that you maintain consistency in your judgement during assessment. Referring to the topic text from time to time will help you maintain judgement consistency.

5. Using the on-line assessment system

There is an on-line relevance assessment system provided at:

http://inex.lip6.fr

which allows you to view the pooled result set of the topics assigned to you for assessment, to browse the IEEE-CS document collection and to record your assessments. Use your username and password to access this system.

After logging in, you will be presented with the Home page (see Figure 2) enlisting the topic ID numbers of the topics assigned to you for assessment (under the title "Choose a pool"). This page can always be reached by clicking on the **Home** link on any subsequent pages.

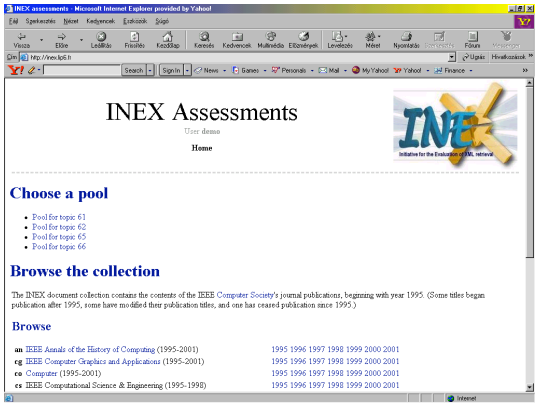


Figure 2. Home page of the assessment system

Clicking on a topic ID will display the pool main page for that topic (see Figure 3).

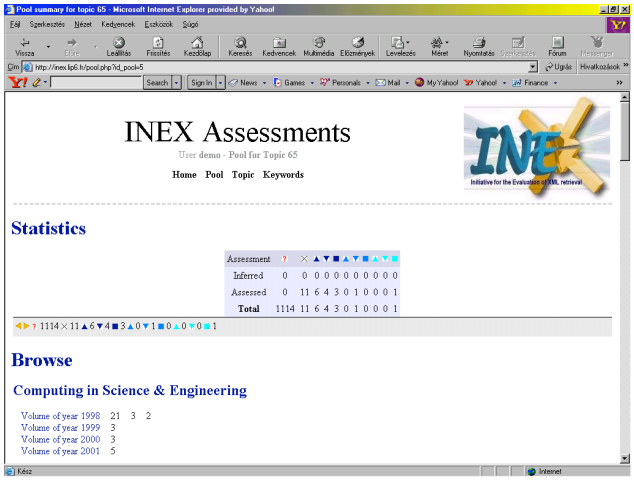


Figure 3. Pool main page

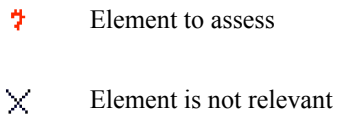
At the top of the pool main page the following links are shown: **Home**, **Pool**, **Topic** and **Keywords**. By clicking on the **Pool** link you can always return to this starting main pool page during your work. By selecting the **Topic** link you can display the topic text in a popup window. This is useful as it allows you to refer to the topic at any time during your assessment. The **Keywords** link allows you to edit a list of *coloured keywords* (cue words or phrases). This feature allows you to specify a list of words or phrases to be highlighted when viewing the contents of an article during assessment. These cue words or phrases can help you in locating potentially relevant texts within an article and will aid you in speeding up your assessment (so add as many relevant cue words as you can think of)! You may edit, add to or delete your list of keywords at any time during your assessment (remember, however, to reload the currently assessed document to reflect the changes). You may also specify the preferred highlighting colour for each and every keyword. After selecting the Keywords link, a popup window will appear showing a table of coloured cells. A border surrounding a cell signifies a colour that is

already used for highlighting some keywords. You can move the mouse cursor over this cell to display the list of keywords that will be highlighted in that colour. To edit the list of words or phrases for a given colour, click on the cell of your choice. You will be prompted to enter a list of words or phrases (one per line) to highlight. Note that the words or phrases you specify will be matched against the text in the assessed documents in their exact form, i.e. no stemming is performed.

In the on-line assessment system, the following scheme is used:

1. *Exhaustivity level* is displayed in different shades of blue.
2. Geometric shapes are used for *specificity level*.

The tables below show the different icons used to indicate the relevance value of an XML element.



Exhaustivity	Highly exhaustive	Fairly exhaustive	Marginally exhaustive
Specificity			
Highly specific			
Fairly specific			
Marginally specific			

Table 1: Icons used to indicate relevance values

Note that all icons except the ? icon can be used by assessors to specify the relevance value (the exhaustivity and specificity level) of an element. The ? icon is used by the on-line assessment system only to mark components that need to be assessed.

This year, the assessment system makes use of two types of inference mechanisms to ensure exhaustive and consistent assessments: we refer to these as passive and active inferences. The passive type simply identifies new elements to be assessed based on those already assessed. For example, for any relevant element (e.g. any component assessed other than "not relevant"), the relevance of its child elements must be assessed, even if these were not part of the original assessment pool (i.e. have not been retrieved). With the application of the passive inference rules, these need-to-be-assessed components will be marked with the ? icon. Unlike the passive rules, the active inference rules are able to derive the relevance value of some elements. These inferred relevance values will be marked using a red border. For example, ☒ denotes "inferred as not relevant", which is assigned to a component if all its child elements have been assessed as "not relevant".

The on-line assessment system provides three main views:

1. The pool view
2. The volume view
3. The article view

In each of these views, a *status bar* appears at the bottom of the window and shows statistics on the current view: how many elements have been assessed as highly exhaustive and highly specific, as highly exhaustive and fairly specific, etc; how many elements have been assessed as not relevant (□); and how many elements remain to be assessed (?). Only when no more elements remain to be assessed is the assessment for that view (pool / volume / article) complete.

In the status bar, three arrows may be used to navigate quickly between the elements to be assessed. The *up arrow* enables you to move from the article view to the volume view or from the volume view to the pool view (you move in the opposite direction by selecting a volume and then an article from the displayed lists). The *left arrow* can be used to go to the previous element to be assessed, while the *right arrow* to go to the next element to be assessed.

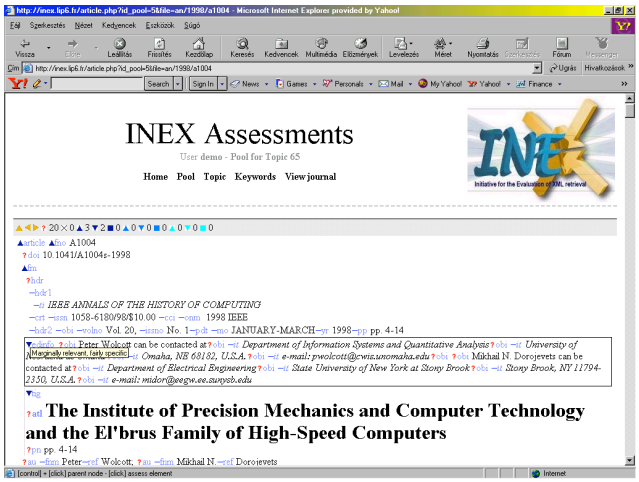


Figure 4. Article view

It is in the article view that elements can be assessed. The article view displays all the elements that form an article, whether these elements are to be assessed or not. In addition, the article view (see Figure 4) shows every XML tag in the article but tries to keep an eye-friendly view of the article. XML tags are displayed between brackets, in light blue, and according to their given (or inferred) assessments when applicable. For instance, an <abs> tag that has been assessed as "highly exhaustive and fairly specific" is displayed as follows:

[▼abs]

The mouse cursor becomes a cross when it is held over an XML tag name. You can then:

- Control-click to scroll to the parent element. The parent element will be highlighted in less than a second (in red).
- Click to display the assessment panel for the element. The assessment panel has three components: the path (first line), the current assessment (second line), and the set of 11 icons (reflecting all possible assignments shown in Table 1). Forbidden assessments (e.g. assessing a parent element as not relevant where one of its child elements is relevant) are displayed in a grey box. To assess the current element, click on the icon with the corresponding relevance value. To hide the panel, click anywhere else in the panel.

Note that you do not need to save your relevance assessments, as the on-line assessment system will automatically do this.

Acknowledgements

A working group was created to discuss the guidelines described in this document. Many thanks go to them for the lively discussion and many inputs. People involved include: Shlomo Geva, Norbert Goevert, Djoerd Hiemstra, Jaana Kekalainen, Shaorong Liu, Anne-Marie Vercoustre, and Arjen de Vries. Also, many thanks to Norbert Goevert for preparing the pools.