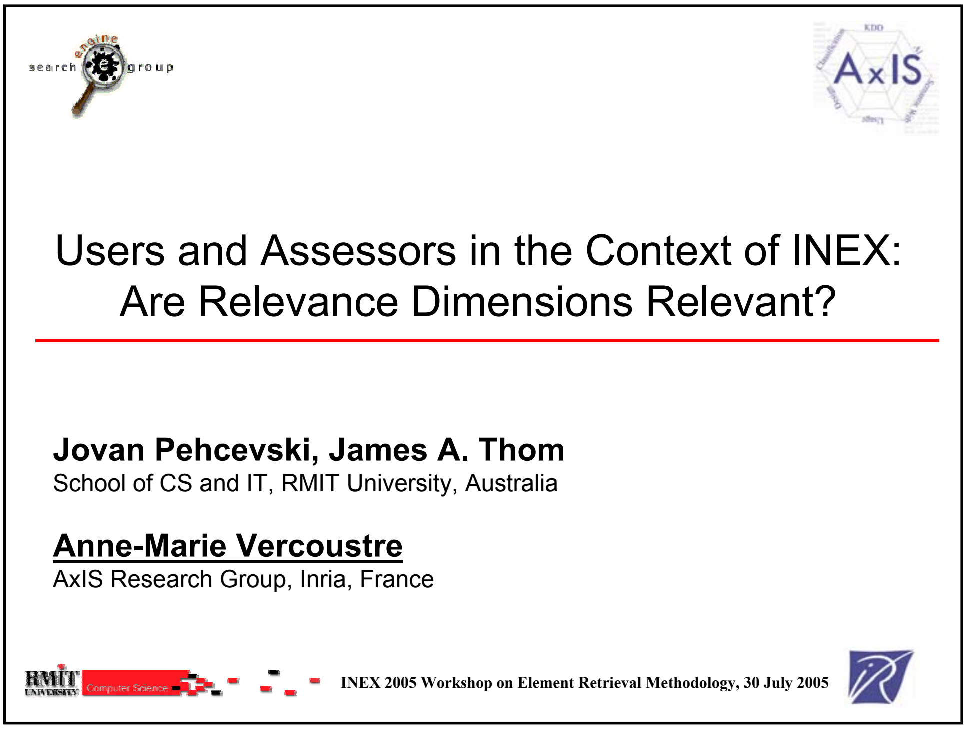
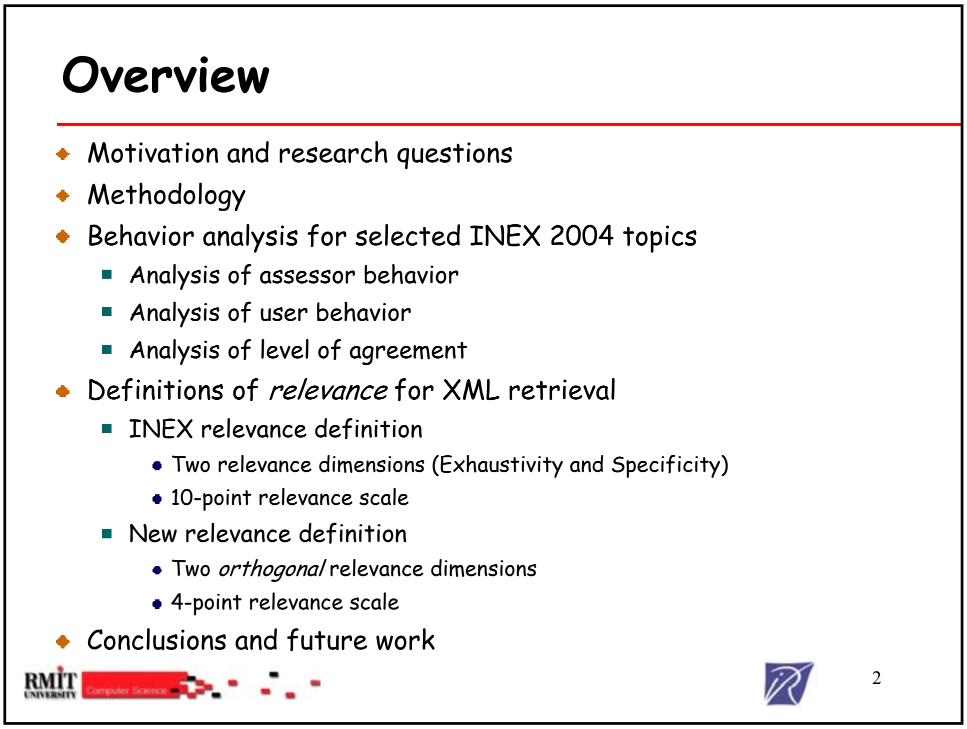


Users and Assessors in the Context of INEX: Are Relevance Dimensions Relevant?

Jovan Pehcevski, James A. Thom
School of CS and IT, RMIT University, Australia

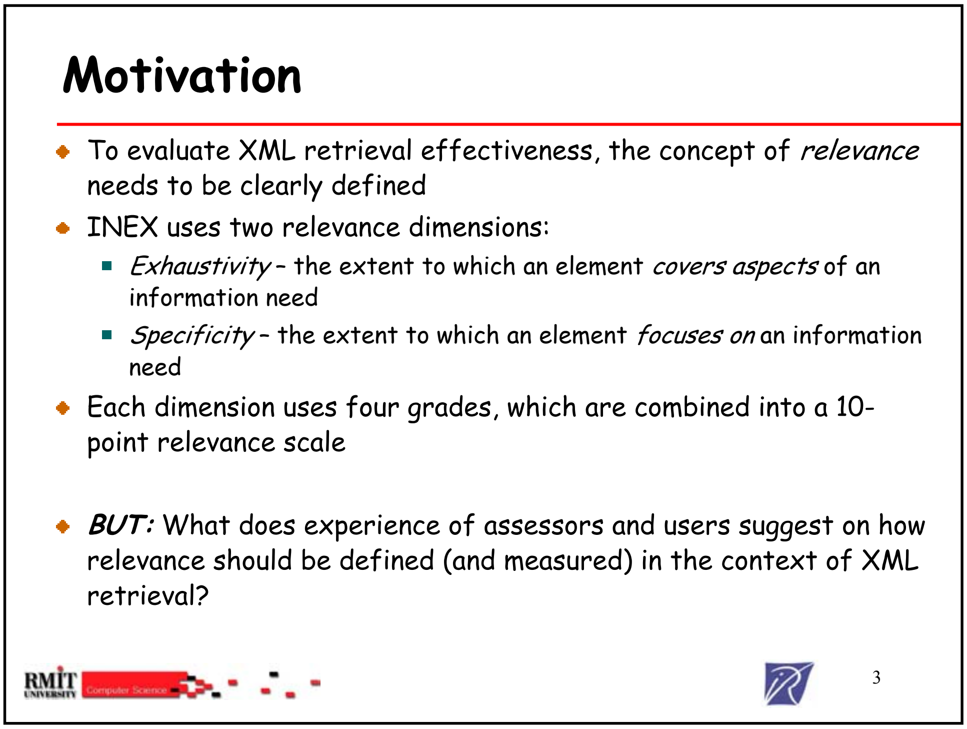
Anne-Marie Vercoustre
AxIS Research Group, Inria, France

INEX 2005 Workshop on Element Retrieval Methodology, 30 July 2005



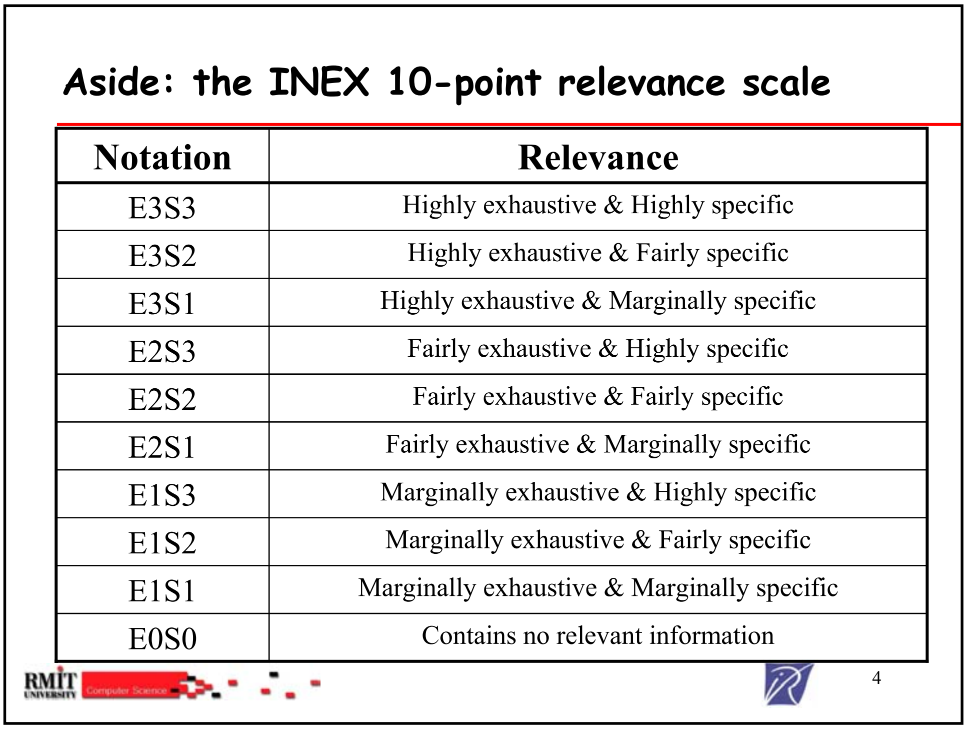
Overview

- Motivation and research questions
- Methodology
- Behavior analysis for selected INEX 2004 topics
  - Analysis of assessor behavior
  - Analysis of user behavior
  - Analysis of level of agreement
- Definitions of *relevance* for XML retrieval
  - INEX relevance definition
    - Two relevance dimensions (Exhaustivity and Specificity)
    - 10-point relevance scale
  - New relevance definition
    - Two *orthogonal* relevance dimensions
    - 4-point relevance scale
- Conclusions and future work



Motivation

- To evaluate XML retrieval effectiveness, the concept of *relevance* needs to be clearly defined
- INEX uses two relevance dimensions:
  - *Exhaustivity* – the extent to which an element *covers aspects* of an information need
  - *Specificity* – the extent to which an element *focuses on* an information need
- Each dimension uses four grades, which are combined into a 10-point relevance scale
- ***BUT:*** What does experience of assessors and users suggest on how relevance should be defined (and measured) in the context of XML retrieval?



Aside: the INEX 10-point relevance scale

Notation	Relevance
E3S3	Highly exhaustive & Highly specific
E3S2	Highly exhaustive & Fairly specific
E3S1	Highly exhaustive & Marginally specific
E2S3	Fairly exhaustive & Highly specific
E2S2	Fairly exhaustive & Fairly specific
E2S1	Fairly exhaustive & Marginally specific
E1S3	Marginally exhaustive & Highly specific
E1S2	Marginally exhaustive & Fairly specific
E1S1	Marginally exhaustive & Marginally specific
E0S0	Contains no relevant information

Research Questions

- Is the INEX 10-point relevance scale *well perceived* by users?
- Is there a *common aspect* influencing the choice of combining the grades of the two INEX relevance dimensions?
- Do users like retrieving *overlapping* document components?
- [Aside: How confusing is for assessors to judge *overlapping* components?]

Methodology used in the study

- Retrieval topics
  - Four INEX 2004 Content Only (CO) topics, reformulated as simulated work task situations (as used in the INEX 2004 Interactive track)
  - Two topic categories: *Background* (topics B1 and B2) and *Comparison* (topics C1 and C2)
- Participants
  - Assessors - topic authors that (in most cases) also assessed the relevance of retrieved elements
  - Users – 88 searchers that participated in the INEX 2004 Interactive track, with no experience in element retrieval
- Collecting the relevance judgments
  - Assessors - judgments obtained from the assessment system
  - Users – judgments obtained from the HyREX log files

Methodology…

- Measuring overlap

  Set-based overlap - for a set of returned elements, the percentage of elements that are *fully contained* by another existing element in the set

- Consider the following set of elements:
  1. /article[1]//sec[1]
  2. /article[1]//sec[1]/ss1[1]
  3. /article[1]//sec[1]/ss1[1]/ss2[1]
  4. /article[1]//sec[2]/ss1[1]
  5. /article[1]//sec[2]
- The set-based overlap is 60%, since three (out of five) elements are fully contained by another element in the set

Methodology…

- Investigating correlation between relevance grades
  - Check whether the choice of combining the grades of the two INEX relevance dimensions is influenced by a common aspect
  - **Sp|Ex (%)** – the percentage of cases where an element is judged as **Sp** (specific), given that it has already been judged to be **Ex** (exhaustive)
  - **Ex|Sp (%)** – the percentage of cases where an element is judged as **Ex** (exhaustive), given that it has already been judged to be **Sp** (specific)

- Consider the following correlation values:
  1. **S3|E3 = 67%** - indicates that in 67% of the cases a highly exhaustive element is also judged to be highly specific
  2. **E2|S3 = 75%** - indicates that in 75% of the cases a highly specific element is also judged to be fairly exhaustive

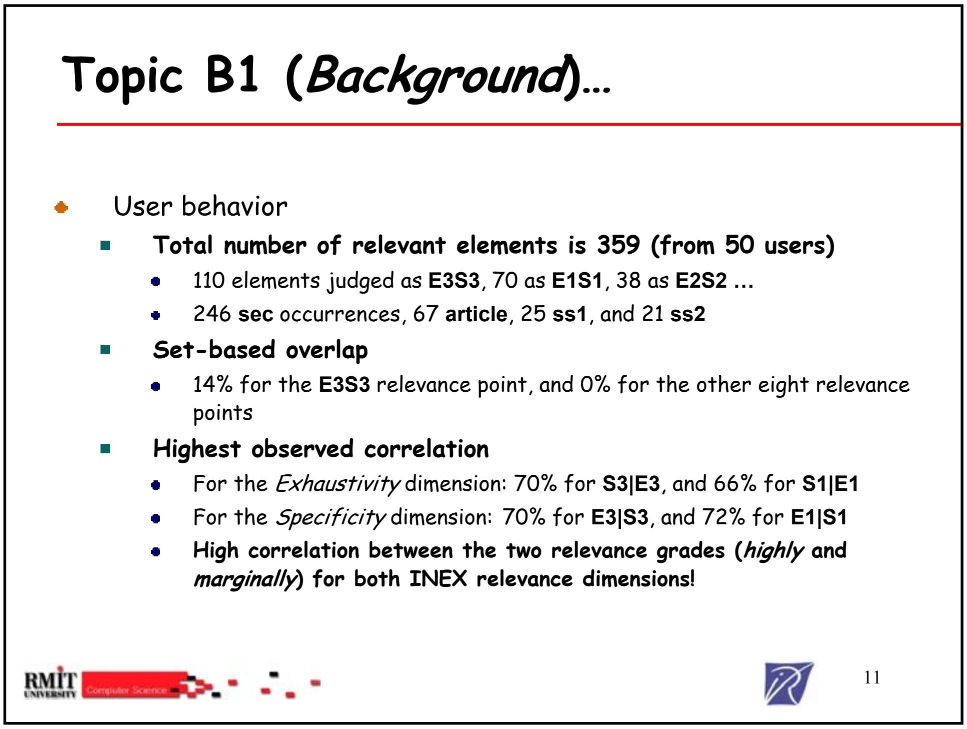
Behavior analysis

- Analysis of assessor behavior
- Analysis of user behavior
- Analysis of level of agreement

- Topics B1 (*Background*) and C2 (*Comparison*) are used in each of the three analysis
  - Relevance judgments collected from around 50 users
  - Assessor judgments available for both topics
- In contrast, for topics B2 and C1
  - Relevance judgments collected from around 18 users
  - No assessor judgments available for topic B2

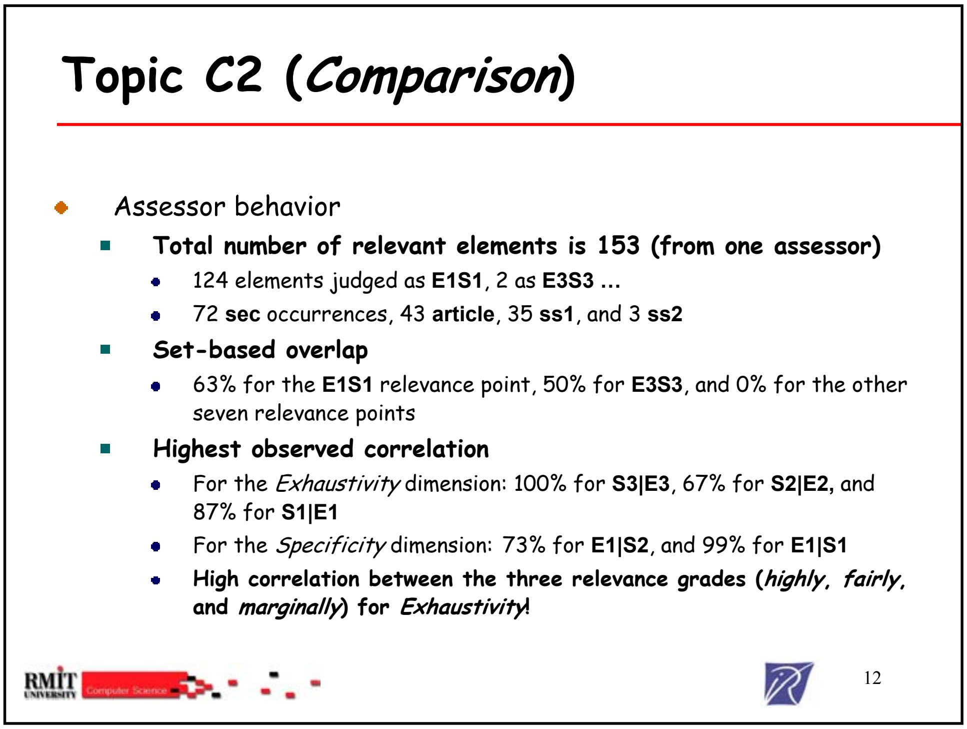
Topic B1 (*Background*)

- Assessor behavior
  - **Total number of relevant elements is 32 (from one assessor)**
    - 11 elements judged as **E2S1**, 9 as **E1S1** …
    - 18 **sec** occurrences, 10 **article**, 3 **ss1**, and one **ss2**
  - **Set-based overlap**
    - 64% for the **E2S1** relevance point, 56% for **E1S1**, and 0% for the other seven relevance points
  - **Highest observed correlation**
    - For the *Exhaustivity* dimension: 67% for **S3|E3**, and 90% for **S1|E1**
    - For the *Specificity* dimension: 75% for **E2|S3**, and 67% for **E2|S2**



Topic B1 (*Background*)…

- User behavior
  - **Total number of relevant elements is 359 (from 50 users)**
    - 110 elements judged as **E3S3**, 70 as **E1S1**, 38 as **E2S2** …
    - 246 **sec** occurrences, 67 **article**, 25 **ss1**, and 21 **ss2**
  - **Set-based overlap**
    - 14% for the **E3S3** relevance point, and 0% for the other eight relevance points
  - **Highest observed correlation**
    - For the *Exhaustivity* dimension: 70% for **S3|E3**, and 66% for **S1|E1**
    - For the *Specificity* dimension: 70% for **E3|S3**, and 72% for **E1|S1**
    - **High correlation between the two relevance grades (*highly* and *marginally*) for both INEX relevance dimensions!**



Topic C2 (*Comparison*)

- Assessor behavior
  - **Total number of relevant elements is 153 (from one assessor)**
    - 124 elements judged as **E1S1**, 2 as **E3S3** …
    - 72 **sec** occurrences, 43 **article**, 35 **ss1**, and 3 **ss2**
  - **Set-based overlap**
    - 63% for the **E1S1** relevance point, 50% for **E3S3**, and 0% for the other seven relevance points
  - **Highest observed correlation**
    - For the *Exhaustivity* dimension: 100% for **S3|E3**, 67% for **S2|E2,** and 87% for **S1|E1**
    - For the *Specificity* dimension: 73% for **E1|S2**, and 99% for **E1|S1**
    - **High correlation between the three relevance grades (*highly*, *fairly*, and *marginally*) for *Exhaustivity*!**

Topic C2 (*Comparison*)…

- User behavior
  - **Total number of relevant elements is 445 (from 52 users)**
    - 101 elements judged as **E1S1**, 66 as **E2S2**, 63 as **E3S3** …
    - 159 **sec** occurrences, 153 **article**, 130 **ss1**, and 3 **ss2**
  - **Set-based overlap**
    - 3% for the **E1S1** relevance point, 9% for **E3S3**, and 0% for the other seven relevance points
  - **Highest observed correlation**
    - For *Exhaustivity*: 53% for **S3|E3**, 44% for **S2|E2**, and 59% for **S1|E1**
    - For *Specificity*: 49% for **E3|S3**, 43% for **E2|S2**, and 64% for **E1|S1**
    - **High correlation between the three relevance grades (*highly*, *fairly*, and *marginally*) for both INEX relevance dimensions!**

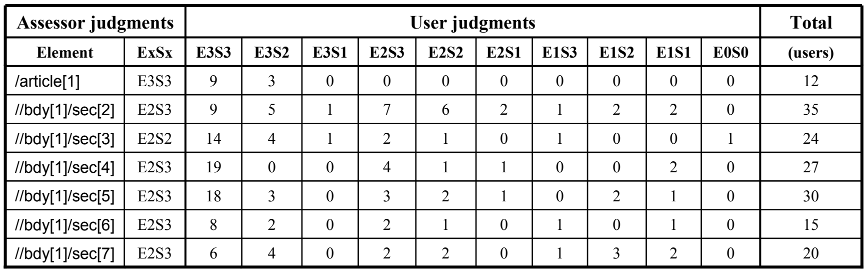
Level of agreement

Topic	Relevance point										Overall (%)
	E3S3 (%)	E3S2 (%)	E3S1 (%)	E2S3 (%)	E2S2 (%)	E2S1 (%)	E1S3 (%)	E1S2 (%)	E1S1 (%)	E0S0 (%)	
B1	52.08	0.00	0.00	14.06	4.17	0.00	0.00	0.00	23.40	56.57	15.10
C2	42.86	0.00	0.00	15.79	15.79	0.00	25.00	16.67	21.74	58.22	19.61

Table 1. Level of agreement between the assessor and the users for each of the topics B1 and C2 (overall and separately for a relevance point).

- The highest level of agreement between the assessor and the users (in both topic cases) is on highly relevant (**E3S3**) and on non-relevant (**E0S0**) elements
- Overall, the agreement for topic C2 is higher than for topic B1

Level of agreement (more detailed)



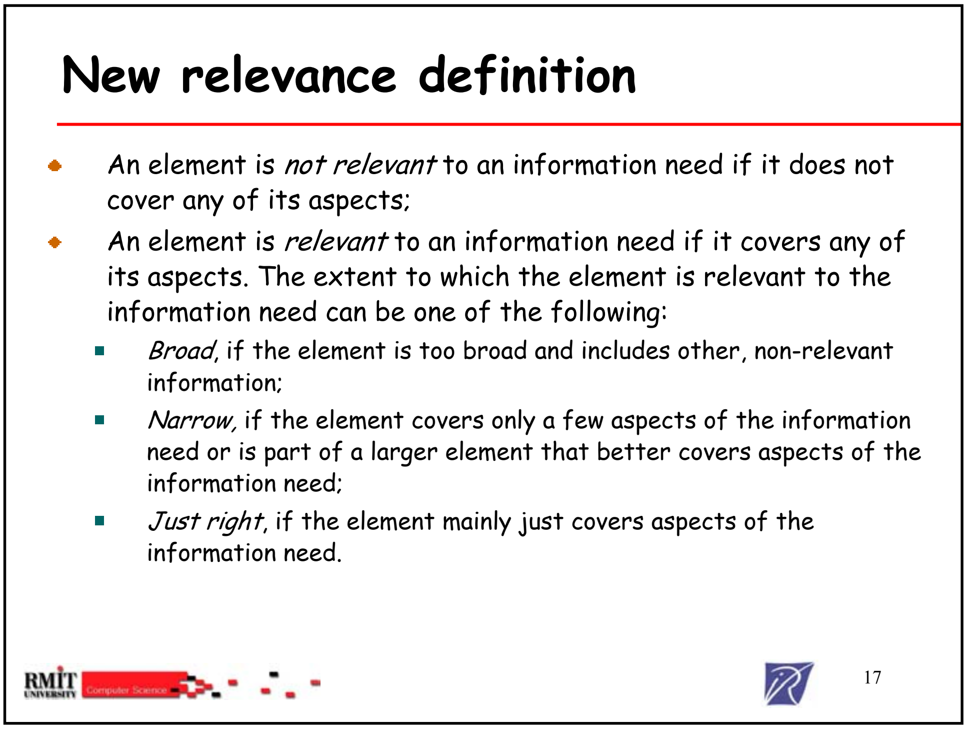
Assessor judgments		User judgments										Total
Element	ExSx	E3S3	E3S2	E3S1	E2S3	E2S2	E2S1	E1S3	E1S2	E1S1	E0S0	(users)
/article[1]	E3S3	9	3	0	0	0	0	0	0	0	0	12
//bdy[1]/sec[2]	E2S3	9	5	1	7	6	2	1	2	2	0	35
//bdy[1]/sec[3]	E2S2	14	4	1	2	1	0	1	0	0	1	24
//bdy[1]/sec[4]	E2S3	19	0	0	4	1	1	0	0	2	0	27
//bdy[1]/sec[5]	E2S3	18	3	0	3	2	1	0	2	1	0	30
//bdy[1]/sec[6]	E2S3	8	2	0	2	1	0	1	0	1	0	15
//bdy[1]/sec[7]	E2S3	6	4	0	2	2	0	1	3	2	0	20

Table 2. Distribution of relevance judgments for the XML file *cg/1998/g1016* (topic B1). For each element judged by the assessor, user judgments and the number of users per judgment are shown.

- The highly relevant element (**article**) was judged by 12 (out of 50) users, and 70% of them also confirmed it to be highly relevant
- ***BUT:*** The **sec[3]** element was judged as **E2S2** (?) by the assessor, while 58% of the users (14 out of 24) judged this element to be **E3S3**!
- Is the *Specificity* dimension misunderstood?

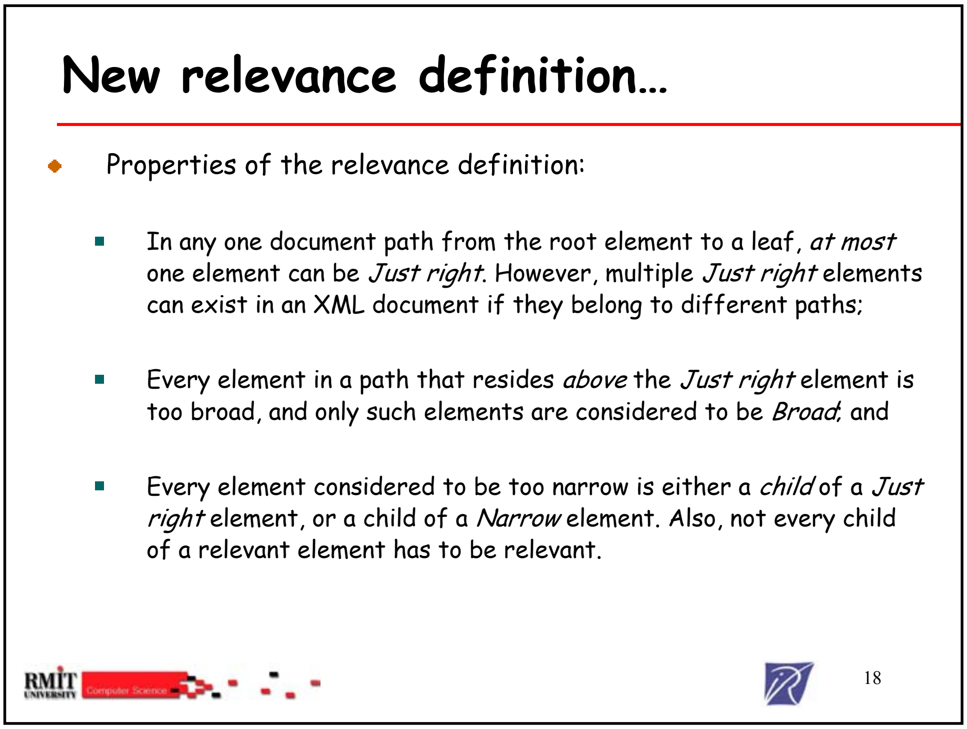
Discussion

- User behavior in the context of XML retrieval
  - There is almost no overlap among relevant elements!
    - **Users do not want to retrieve redundant information?**
  - Highest observed correlation exists between the same grades of the two INEX relevance dimensions!
    - **The cognitive load of simultaneously choosing the grades for *Exhaustivity* and *Specificity* is too difficult a task?**
- Level of agreement between the assessor and the users
  - The highest level of agreement is on the end points of the INEX 10-point relevance scale!
    - **Only the end points of the relevance scale are perceived in the same way by both the assessor and the users?**
- **Perhaps a simpler relevance definition is needed for INEX?**



New relevance definition

- An element is *not relevant* to an information need if it does not cover any of its aspects;
- An element is *relevant* to an information need if it covers any of its aspects. The extent to which the element is relevant to the information need can be one of the following:
  - *Broad*, if the element is too broad and includes other, non-relevant information;
  - *Narrow,* if the element covers only a few aspects of the information need or is part of a larger element that better covers aspects of the information need;
  - *Just right*, if the element mainly just covers aspects of the information need.



New relevance definition…

- Properties of the relevance definition:
  - In any one document path from the root element to a leaf, *at most* one element can be *Just right*. However, multiple *Just right* elements can exist in an XML document if they belong to different paths;
  - Every element in a path that resides *above* the *Just right* element is too broad, and only such elements are considered to be *Broad*; and
  - Every element considered to be too narrow is either a *child* of a *Just right* element, or a child of a *Narrow* element. Also, not every child of a relevant element has to be relevant.

New relevance definition…

- Partial mapping to the INEX relevance scale:
  - *Non-relevant* <=> E = 0, S = 0 (**E0S0**)
  - *Just right* <=> E = 3, S = 3 (**E3S3**)
  - *Broad* <=> E = 3, S < 3 (**E3S2, E3S1**)
  - *Narrow* <=> E < 3, S = 3 (**E2S3, E1S3**)

New relevance definition…

- Much simpler compared to the current INEX definition
  - Two *orthogonal* relevance dimensions: first based on topical relevance (with a binary scale); second based on hierarchical relationships among elements (with three relevance grades)
  - 4-point relevance scale (instead of 10-point relevance scale)
  - Should reduce the cognitive load of assessors and users
- Allows to explore different aspects of XML retrieval
  - Are different retrieval techniques needed to retrieve *Just right*, rather than any *Broad* or *Narrow*, relevant elements?
- Requires almost no need to modify some of the INEX metrics (?)
- Works well with the yellow highlighter ☺

Conclusions and future work

- Is the INEX 10-point relevance scale *well perceived* by users?
  - Only the end points of the relevance scale are well perceived
- Is there a *common aspect* influencing the choice of combining the grades of the two INEX relevance dimensions?
  - Users behave as if each of the grades from either dimension belongs to only one relevance dimension, suggesting that the cognitive load of simultaneously choosing the relevance grades is too difficult a task
- Do users like retrieving *overlapping* document components?
  - Users do not want to retrieve, and thus do not tolerate, redundant information
- Analysis of a greater number of topics will be undertaken in the future to confirm the significance of the above findings

Questions?

The Twelve Apostles, Port Cambell National Park

The church of St. Jovan the Divine at Kaneo, Ohrid

In Australia !

In Macedonia !

(Greetings from Jovan and Jamie) ☺