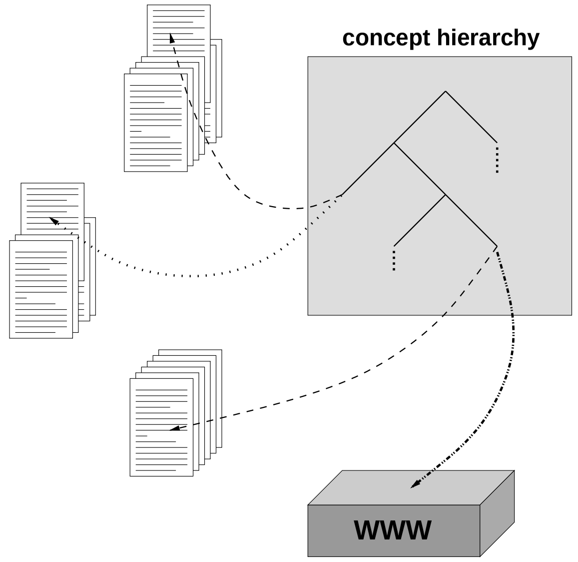
Concept-Based Computer-Aided Link Generation for Electronic Handbooks

Christof Monz Jon Ragetli Maarten de Rijke
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
{christof, ragetli, mdr}@science.uva.nl
www.illc.uva.nl/~{christof, ragetli, mdr}

overview

- collaboration between Elsevier Science Publisher and the University of Amsterdam
- the aim of the project is to ...

architecture



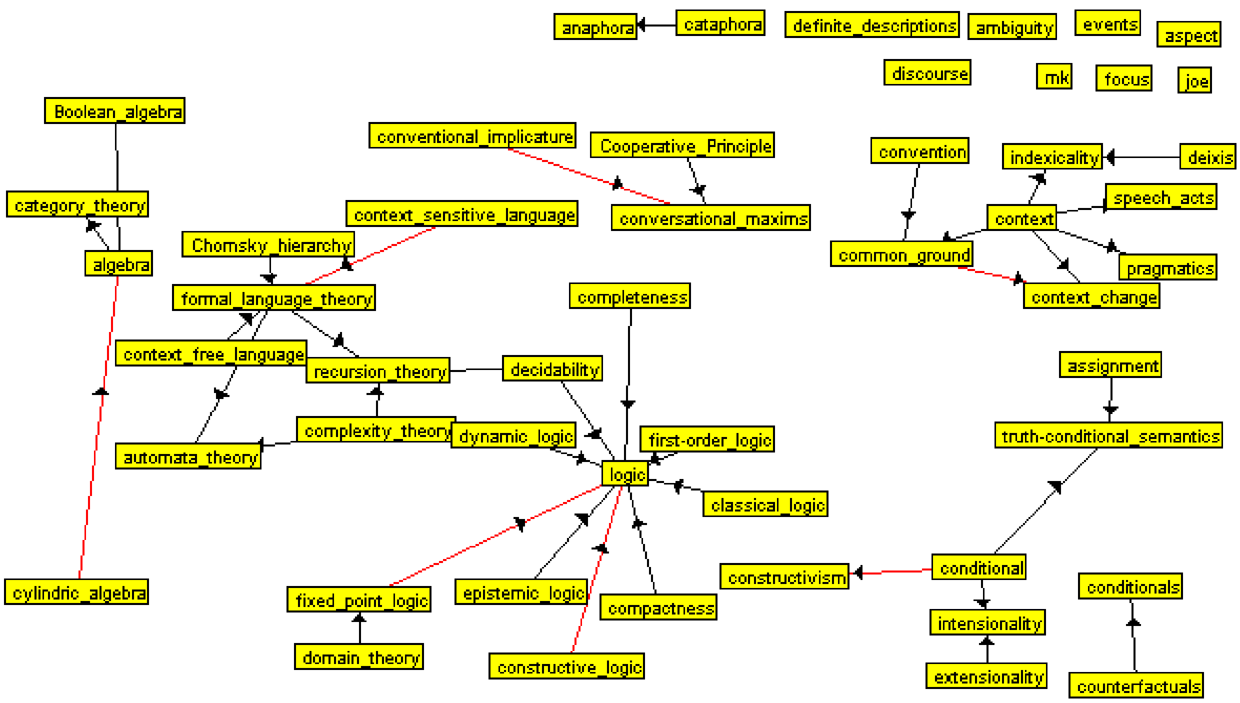
concept hierarchy

- nodes in the hierarchy are terms from the handbook's glossary together with their explanations
- each head of a glossary entry is augmented by a list of synonymous terms
- currently, the concept hierarchy contains around 500 nodes

building a concept hierarchy

- manually
- automatically

excerpt of the concept hierarchy



linking

- types of links
- automatically creating links

types of links

currently there are four types of links:

- definition: definition of a particular glossary entry
- result: important theorems or propositions related to an entry
- related text: passages that are semantically related to an entry
- references: bibliographic entries related to the glossary entry

automated linking

- links to definitions and results employ pattern matching to identify the relevant portions of text
  - a glossary entry is linked to a definition or result if its head appears in it
- relevant text is identified by computing a similarity score between a glossary entry and a text passage
  - similarity is computed as the inner product between the glossary entry and the text passage
- related references are linked by identifying relevant text passages and looking for citations within those passages

applying IR techniques

a passage p is linked to a glossary entry g if the similarity score $sim(g,p)$ is

- above a certain threshold τ
- p is among the n top-ranked passages

experimental issues

- what is an appropriate definition of $sim(g, p)$
- what should count as a passage
- what should be included in the representations of g and p
- what is an optimal value for τ
- how many top-ranked passages should be considered

preliminary setup

- $sim(g, p)$ is defined as the inner product of g and p where g is split into two sub-vectors g_h (glossary head) and g_e (glossary explanation):

$$sim(g, p) = \alpha \cdot (p \bullet g_h) + p \bullet g_e$$

- term weights are computed using Salton & Buckleys's $tfx.nfc$ weighting scheme
- a passage is a fixed-length unit of text (350 words)
- only words (no phrases) are indexed

evaluation of the links

problems when evaluating automatically constructed links:

- no standard reference test collections
- inter-user disagreement on relevancy of a link

in this project evaluation was done by

- arbitrarily selecting glossary entries (≈ 30) and manually assessing their top n related passages (links)
- quality is purely defined in terms of precision

conclusions

- blurb