A Conceptual Representation of Documents and Queries for Information Retrieval Based on Light Ontologies

Andrea G. B. Tettamanzi[1]

Università degli Studi di Milano, Dipartimento di Tecnologie dell'Informazione
Via Bramante 65, 26013 Crema (CR), Italy
andrea.tettamanzi@unimi.it

Sophia Antipolis, Tuesday April 3, 2012

[1] Joint work with Célia da Costa Pereira and Mauro Dragoni.

From Query Expansion to Document Semantic Expansion

- Expansion techniques are generally related to queries.
  - by using thesauri (manual or automatic);
  - by adding, to queries, terms that are synonyms or related to the term to expand;
- Only recently[2], expansion has been applied to documents.
  - idea: documents and queries are represented in the same way;
  - the importance of how many and which terms have to be used for expansion decreases;
  - however, this kind of approach presents an issue related to term coverage;

[2] M. Baziz, M. Boughanem, G. Pasi, and H. Prade, "An Information Retrieval Driven by Ontology: from Query to Document Expansion", RIAO 2007

The Intuition Behind

Starting point:

Considering how information is usually represented and classified.

Issues:

Drawbacks of the term-based representation.

Challenge:

Using concepts to represent terms in documents and queries.

IMPORTANT:

This is not a classic expansion technique!

Roadmap to a Concept-Based Representation

- Choose a method allowing to represent all document and query terms by using the same set of concepts.
- Assign an appropriate weight to each concept, in both documents and queries.

What is a Concept Occurrence?

Concepts do not occur as such in documents!

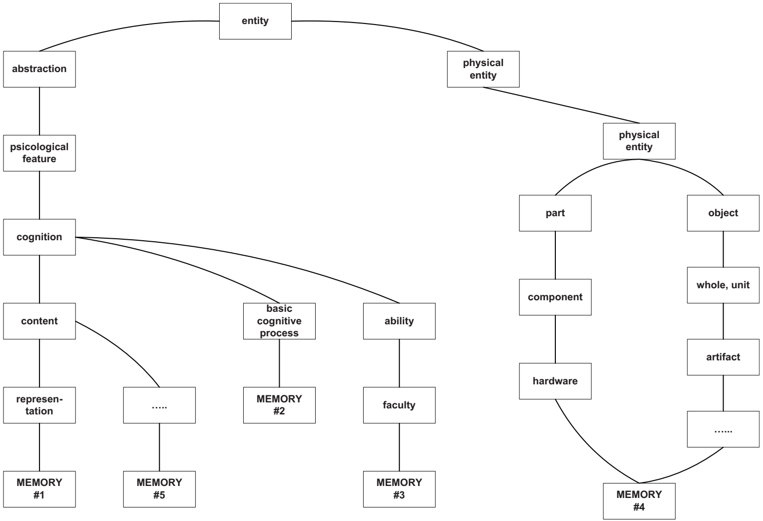
Concept Occurrence

- Concepts occur throgh their *lexicalizations*
- Each term (= word, phrase) may correspond to one or more concepts
- Moreover, a concept may occur implicitly, through any of its *super-* and *subconcepts*

Idea:

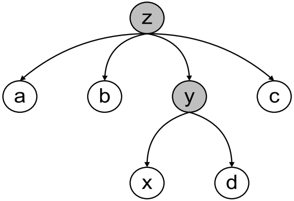
1. Impute a term occurrence to *all* of its senses
2. Distribute concept occurrences over the `is-a` hierarchy

Example: the Word “MEMORY”



Choosing Concepts

- Use WordNet as a "light" ontology;
- Consider a set of independent concepts (⇒ base vector);
- Example: assume we have the following ontology:

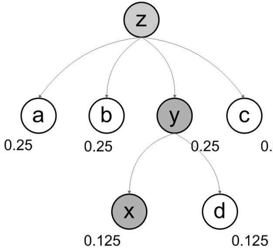


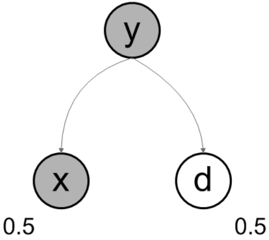
- In this case, the base-vector is: $I = \{a, b, c, d, x\}$

Computing Weights (1)

For each concept, compute explicit and implicit occurrences:

$$N(c) = \mathrm{occ}(c) + \sum_{c \in \mathrm{Path}(c,\ldots,\top)} \sum_{i=2}^{\mathrm{depth}(c)} \frac{\mathrm{occ}(c_i)}{\prod_{j=2}^{i} \|\mathrm{children}(c_j)\|}$$





Computing Weights (2)

Compute the information vector for each concept:

$$
\begin{aligned}
\mathrm{info}(z) &= (0.25, 0.25, 0.25, 0.125, 0.125) \\
\mathrm{info}(a) &= (1.0, 0.0, 0.0, 0.0, 0.0) \\
\mathrm{info}(b) &= (0.0, 1.0, 0.0, 0.0, 0.0) \\
\mathrm{info}(c) &= (0.0, 0.0, 1.0, 0.0, 0.0) \\
\mathrm{info}(y) &= (0.0, 0.0, 0.0, 0.5, 0.5) \\
\mathrm{info}(d) &= (0.0, 0.0, 0.0, 1.0, 0.0) \\
\mathrm{info}(x) &= (0.0, 0.0, 0.0, 0.0, 1.0)
\end{aligned}
$$

To encode document $D =$ “*xxyyyz*”, sum the information vectors of the concepts occurring in it:

$$\mathbf{d} = 2 \cdot \mathrm{info}(x) + 3 \cdot \mathrm{info}(y) + \mathrm{info}(z) = (0.25, 0.25, 0.25, 1.625, 3.625)$$

Implementation

- On top of the Apache Lucene open-source API
- In the pre-indexing phase, documents converted to conceptual representation
- Discard concepts with weight < 0.01
- Concept weights are stored as “payloads”
- Queries converted to conceptual representation as well
- Concept weights are stored as “boost values”

Comparison between Term-Based and Concept-Based Representation

Collection	Number of Documents	Term-Based	
		Vector Size (# of tokens)	Index Size
MuchMore	7823	47623	∼ 3Mbyte
TREC Ad-Hoc	528155	650160	∼ 2Gbyte

Collection	Number of Documents	Concept-Based	
		Vector Size (# of tokens)	Index Size
MuchMore	7823	57708	∼ 5Mbyte
TREC Ad-Hoc	528155	57708	∼ 3.2Gbyte

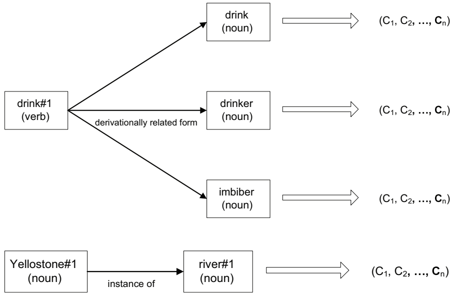
Collection	Number of Documents	Difference	
		Vector Size	Index Size
MuchMore	7823	+ 21.18 %	+ 66.67 %
TREC Ad-Hoc	528155	- 91.12 %	+ 60.00 %

Verbs, Adjectives, and Proper Nouns

Problem: `is-a` relation defined in WordNet for common nouns only

Workaround:

Exploit the "derivationally related form" and the "instance of" relations.



Experimental Protocol

Two Phases:

1. comparison to the most well-known state-of-the-art semantic expansion techniques:
    - document representation by synsets
    - document representation by semantic trees
2. validation with systems that use semantic expansion presented at the TREC7 and TREC8 conferences.

The evaluation method follows the TREC protocol. For each query, the first 1,000 documents have been retrieved and the precision calculated at 5, 10, 15, and 30 documents retrieved.

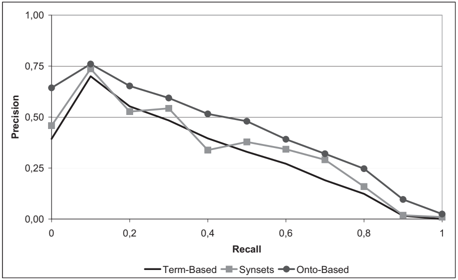
First Phase

- MuchMore Collection (7,823 Documents and 25 Queries)
- Documents of the Springer corpus of parallel medical scientific abstracts
- Relevance assessments provided for each query
- URL: `http://muchmore.dfki.de`

The Term-Based representation has an advantage here, due to the absence of specific medical-domain terms in WordNet

Results on the MuchMore Collection

Precision/recall Graph:



Precision@X and MAP Values:

Systems	Precisions				
	P5	P10	P15	P30	MAP
Term Indexing (Baseline)	0.544	0.480	0.405	0.273	0.449
Synset Indexing by Gonzalo et al. (1998)	0.648	0.484	0.403	0.309	0.459
Conceptual Indexing by Baziz et al. (2007)	0.770	0.735	0.690	0.523	0.449
Proposed Approach	**0.784**	**0.765**	**0.728**	**0.594**	**0.477**

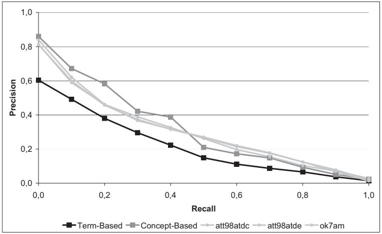
Second Phase

- TREC Ad-Hoc Collection Volumes 4 and 5 (containing over 500,000 documents)
- The approach has been evaluated on topics from 351 to 450
- These topics correspond to TREC-7 and TREC-8
- The index contains documents from
  - Financial Times Ltd. (1991, 1992, 1993, 1994)
  - Congressional Record of the 103rd Congress (1993)
  - Foreign Broadcast Information Service (1996)
  - Los Angeles Times (1989, 1990)
- Comparison to 3 systems presented at TREC-7 and TREC-8
  - based on semantic expansion
  - with the highest precision at low recall levels

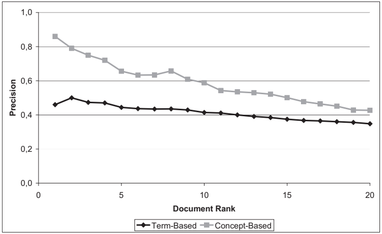
Note that 89% of search result click activity occurs on the 1st page! This means on the 10–20 top-ranking documents.

Results on the TREC-7 Ad-Hoc Collection

Precision/recall Graph:



Precision@20 Graph:

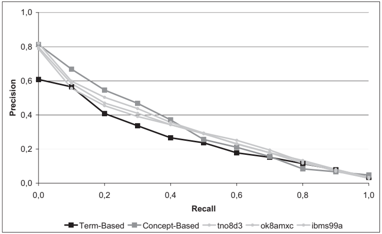


Precision@X and MAP Values:

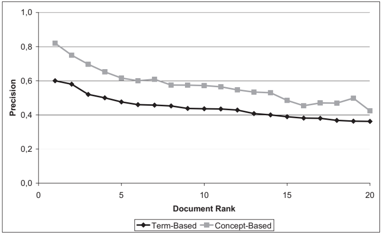
Systems	Precisions				
	P5	P10	P15	P30	MAP
Term-Based Representation	0.444	0.414	0.375	0.348	0.199
AT&T Labs Research (att98atdc)	0.644	0.558	0.499	0.419	0.296
AT&T Labs Research (att98atde)	0.644	0.558	0.497	0.413	0.294
City University, Univ. of Sheffield, Microsoft (ok7am)	0.572	0.542	**0.507**	0.412	0.288
Proposed Approach	**0.656**	**0.588**	0.501	0.397	**0.309**

Results on the TREC-8 Ad-Hoc Collection

Precision/recall Graph:



Precision@20 Graph:



Precision@X and MAP Values:

Systems	Precisions				
	P5	P10	P15	P30	MAP
Term-Based Representation	0.476	0.436	0.389	0.362	0.243
IBM T.J. Watson Research Center (ibms99a)	0.588	0.504	0.472	0.410	0.301
Microsoft Research Ltd (ok8amxc)	0.580	0.550	**0.499**	**0.425**	**0.317**
TwentyOne (tno8d3)	0.500	0.454	0.433	0.368	0.292
Proposed Approach	**0.616**	**0.572**	0.485	0.415	0.315

Tests of Significance

Significance Levels for the Hypotheses:

“the Proposed Approach is Better (Worse) than the Best Benchmark”:

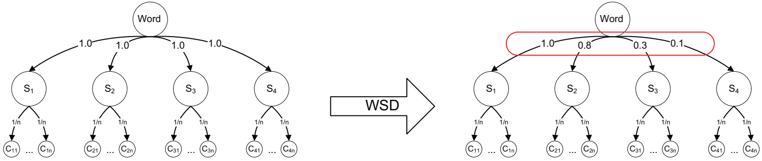
	P5 (%)	P10 (%)	P15 (%)	P30 (%)	MAP (%)
MuchMore	70.72	96.84	99.06	99.99	74.66
TREC-7	57.19	94.39	(29.56)	(84.14)	63.21
TREC-8	92.79	83.80	(62.40)	(47.76)	(10.81)

Possible Research Directions

- The absence of some terms in the ontology, (in particular terms related to specific domains like biomedical, mechanical, business, etc.), may negatively affect the performance of retrieval
- The way proper names are treated is still too simplistic
- Term ambiguity: using a WSD approach would be an improvement
- The proposed approach to representation may be extended
  - Beyond Information Retrieval
  - Beyond Document/Query Representation

Using Word Sense Disambiguation

Assume a WSD system is available, which outputs degrees of possibility/likelihood for each sense, for an occurrence of a polysemous term.



A. Azzini, C. da Costa Pereira, M. Dragoni, and A. G. B. Tettamanzi. "A Neuro Evolutionary Corpus-based Method for Word Sense Disambiguation". *IEEE Intelligent Systems*, in press. DOI: 10.1109/MIS.2011.108

Beyond Information Retrieval

"Conceptual" Folksonomies:

- **resources** are described by bags of tags
- treat **tags** as terms
- use an automatically constructed **tag ontology** instead of WordNet
- map a bag of tags to a resource **concept vector**
- use conceptual representation to enhance ontology construction
- use in a recommender system to compute similarity between users

Relevant Publications

- C. da Costa Pereira and A. G. B. Tettamanzi. “An Ontology-Based Method for User Model Acquisition. In Z. Ma (Ed.), *Soft Computing in Ontologies and Semantic Web*, p. 211-227, Springer, 2006.
  DOI: 10.1007/978-3-540-33473-6_8
- M. Dragoni, C. da Costa Pereira, and A. G. B. Tettamanzi. “An Ontological Representation of Documents and Queries for Information Retrieval Systems”. IEA/AIE 2010.
- M. Dragoni, C. da Costa Pereira, and A. G. B. Tettamanzi. “A Conceptual Representation of Documents and Queries for Information Retrieval Systems by Using Light Ontologies”. *Expert Systems with Applications*, in press.
  DOI: 10.1016/j.eswa.2012.01.188