Automatic Query Expansion Using Random Indexing

Magnus Sahlgren, Jussi Karlgren, Rickard Cöster

Timo Järvinen

Random Indexing

- Vector-space models:
  - Representing words as *context vectors* in multi-dimensional space
  - The context vectors are constructed from co-occurrence statistics
- Random Indexing:
  - *Distributed* representations
  - Efficient, flexible, scalable and robust

Random Indexing

- Assign an 1 800-dimensional sparse random *index vector* to each word:

  filmen = 0000000+0000000000000000…

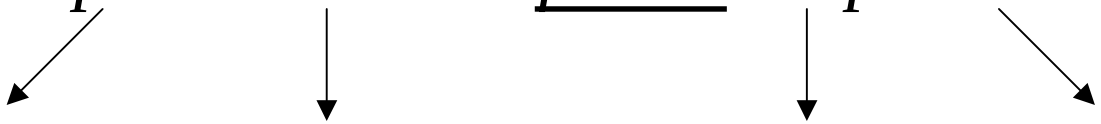
  Pulp = 000000000000000000-00000…

  Fiction = 00000-000000000000000+0…

Random Indexing

- Every time a word occurs in the data, add the index vectors of the *n* surrounding words to the context vector for the word

"…pris tilldelades filmen Pulp Fiction…"



{000-00…} + {0000+0…} + {0+0000…} + {0-0000…}

=

filmen = {0,+1,-4,0,0,+12,+3,0,-34,+1,0,-1,+5,0,-9,+33…}

The Rationale of Context Vectors

- Represent the distributional patterns of words
- Distributional similarity may be expressed in terms of vector similarity
- The *Distributional Hypothesis*:

  Two distributionally similar words have similar meanings

The Rationale of Context Vectors

- Extract semantically similar words:
  - Generate (domain specific) thesauri
  - Automatic query expansion

The CLEF 2002 Experiments

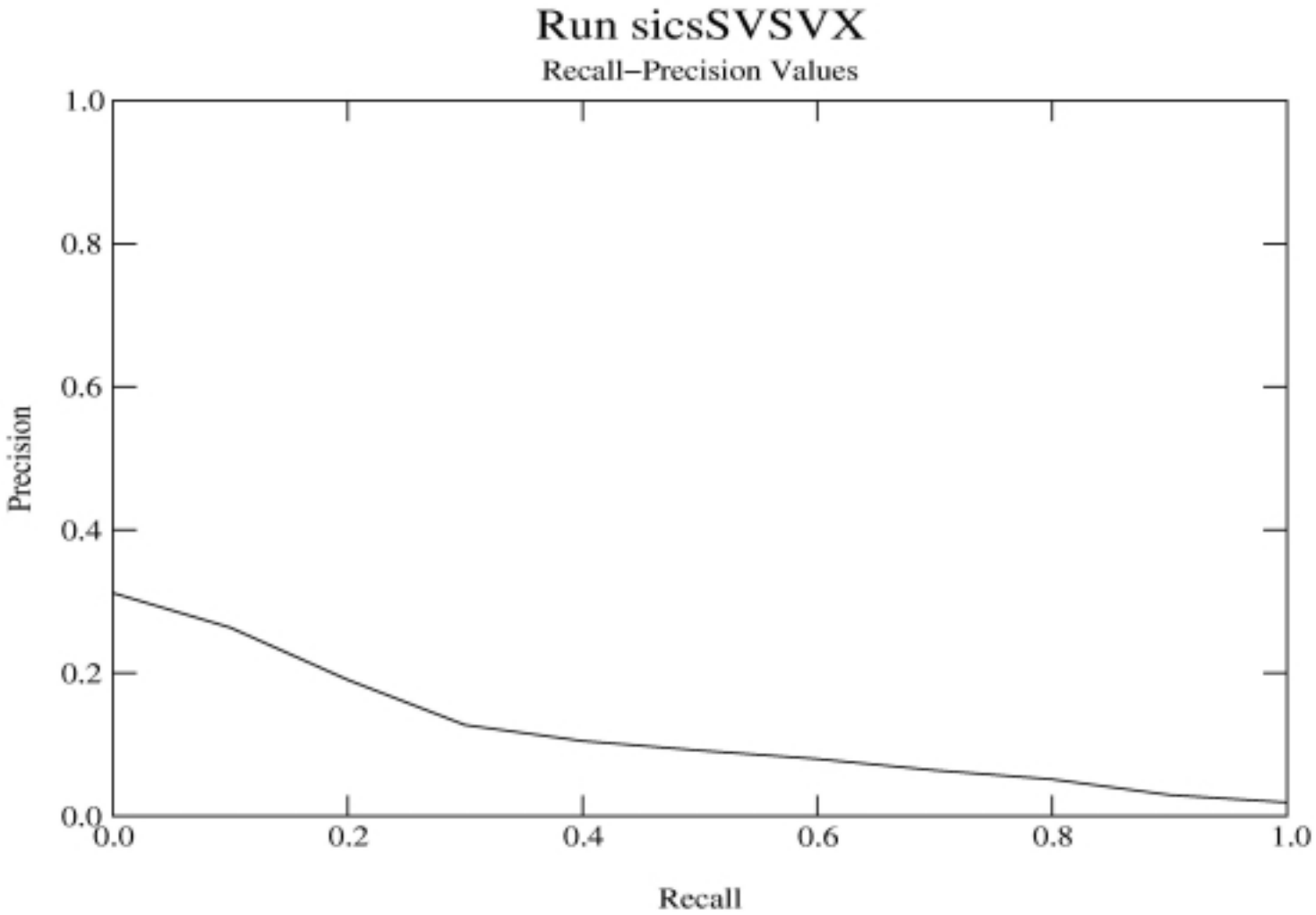
- Conexor's FDG parser normalizes the data to base forms
- Apply Random Indexing:
  - 1 800-dimensional vectors
  - 8 non-zero elements (four +1s and four –1s)
  - 3 + 3 sized context window:

    [0,25 0,5 1] *word* [1 0,5 0,25]

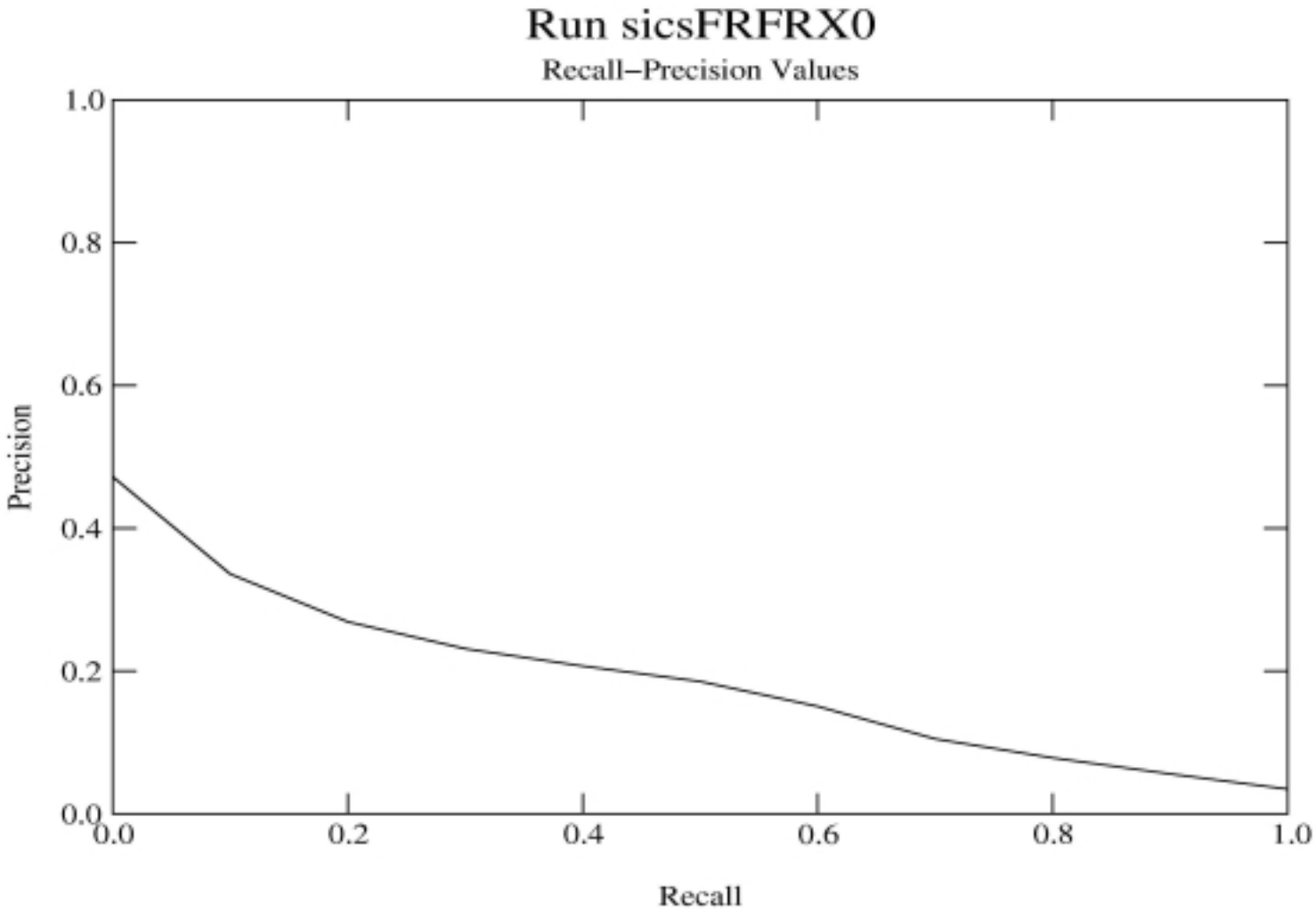
Automatic Query Expansion

- Standard query:
  - Conexor's parser to extract base forms
  - Remove stop words and some query specific terminology
  - Title and description fields
- Expanded query:
  - For every word, add the 5 most similar words

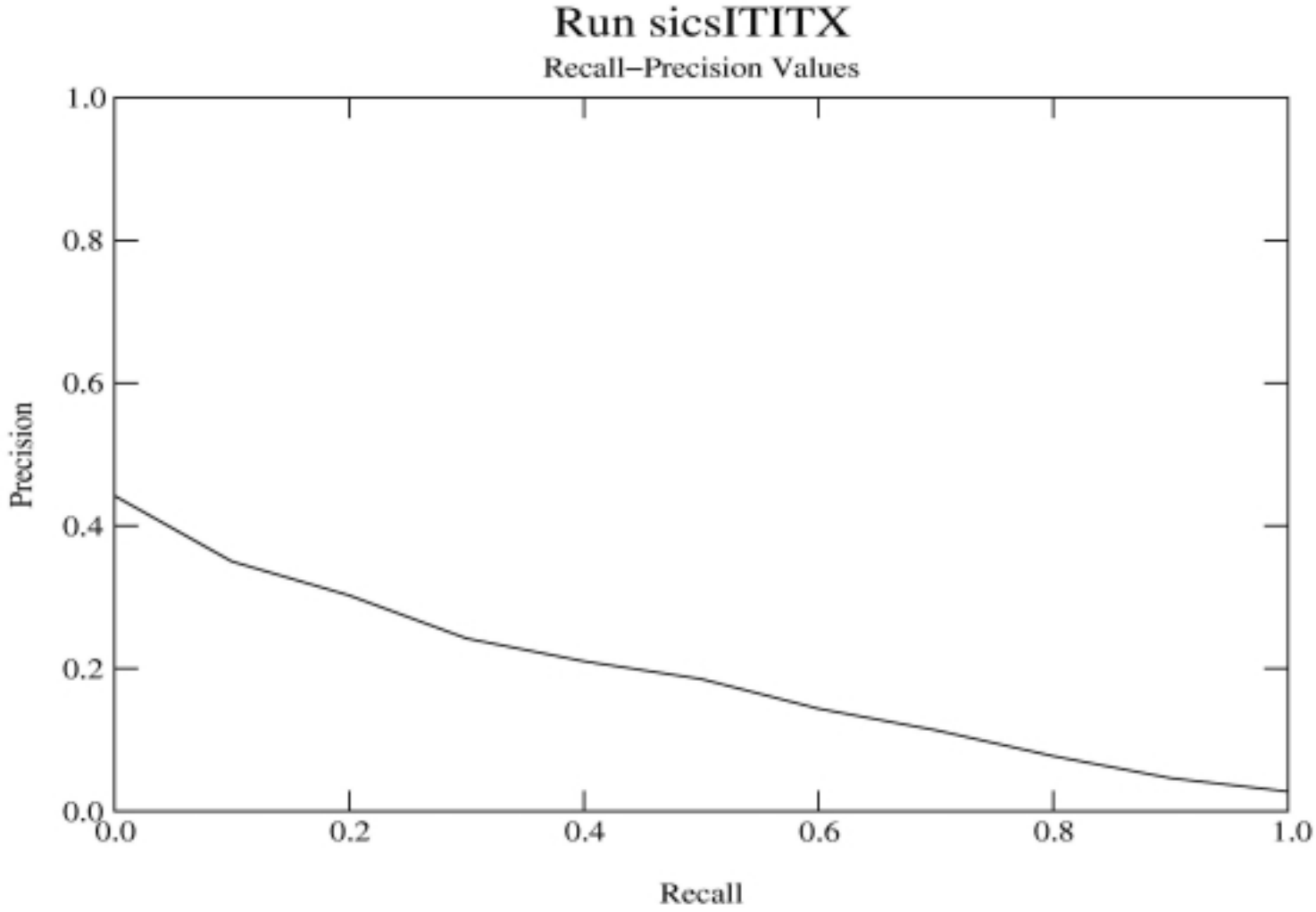
Swedish Results



French Results



Italian Results



Summary of Results

- Baseline results (unexpanded queries) unsatisfactory
- Expanded queries lower the results

Query Construction

- Which words are relevant?
- Query term weighting
- Utilize clause-internal dependencies

Lexical Factors

- Which words should be expanded?
  - Person names?

    Marie → Claude Pierre Gabin Harlow Francios
  - Place names?

    Finland → Norge Danmark Sverige Österrike Island
- Lexical categories

Selection of Expansion Terms

- Word-based expansion:
  - Expand every word in the query
- Concept-based expansion (Qiu & Frei, 1993):
  - Expand a query vector (produced e.g. by summing the context vectors of the words in the query)

Thank you!

Magnus Sahlgren, SICS
CLEF 2002