Syntax versus Semantics:
Analysis of Enriched Vector Space Models

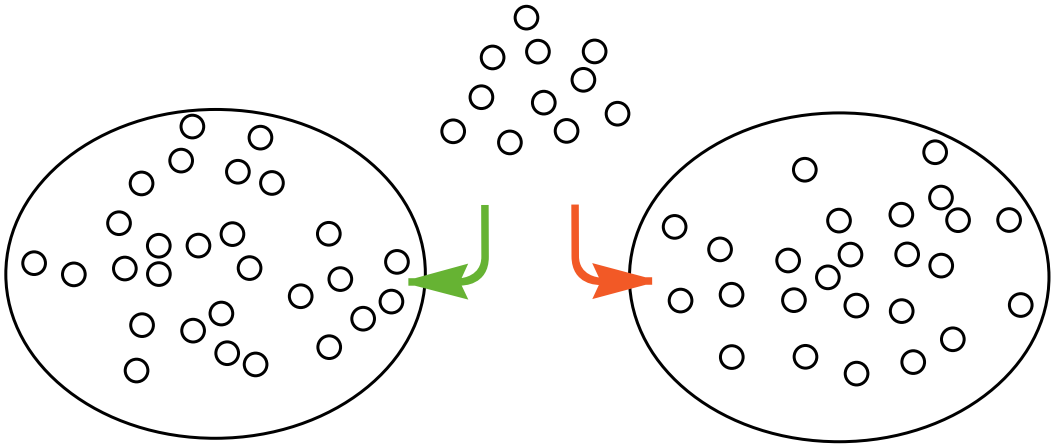
Benno Stein and Sven Meyer zu Eissen and Martin Potthast
Bauhaus University Weimar

Relevance Computation

Information retrieval aims at dividing relevant documents from irrelevant ones with respect to an information need.

Document models are at the heart of such a process.

A look behind the scenes:

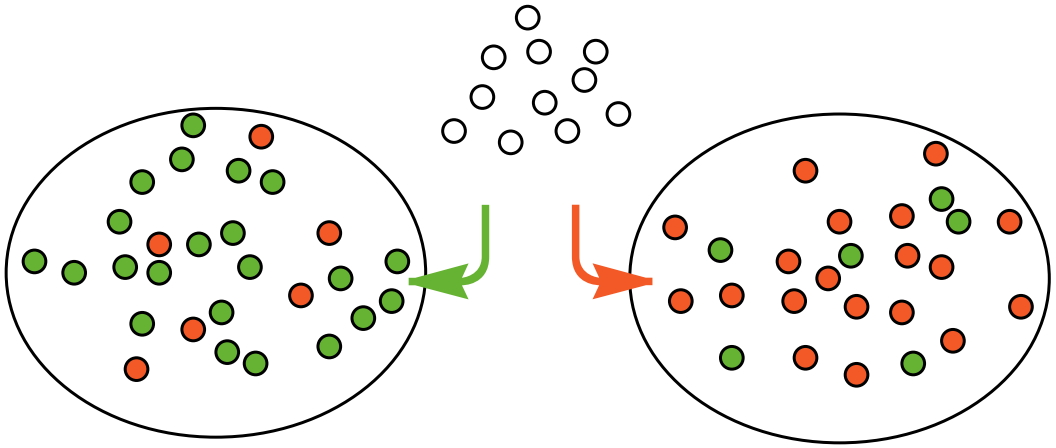


Relevance Computation

Information retrieval aims at dividing relevant documents from irrelevant ones with respect to an information need.

Document models are at the heart of such a process.

A look behind the scenes: An average document model.

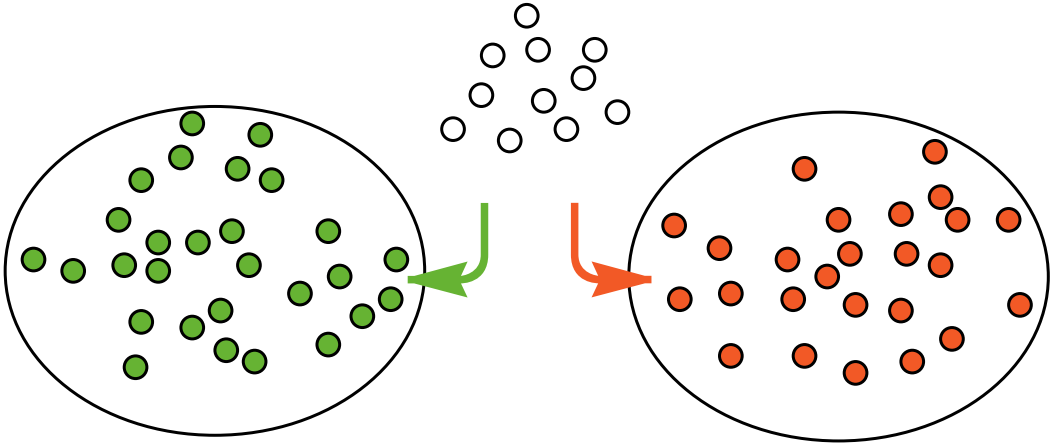


Relevance Computation

Information retrieval aims at dividing relevant documents from irrelevant ones with respect to an information need.

Document models are at the heart of such a process.

A look behind the scenes: A perfect document model.



Index Construction

Text with markups [Reuters]:

```
<TEXT> <TITLE>CHRYSLER> DEAL LEAVES UNCERTAINTY
FOR AMC WORKERS</TITLE> <AUTHOR> By Richard
Walker, Reuters</AUTHOR> <DATELINE> DETROIT,
March 11 - </DATELINE><BODY>Chrysler Corp's 1.5
billion dlr bid to takeover American Motors Corp;
AMO> should help bolster the small automaker's
sales, but it leaves the future of its 19,000
employees in doubt, industry analysts say.  It
was "business as usual" yesterday at the American
...
```

Index construction

Raw text:

```
chrysler deal leaves uncertainty for amc workers
by richard walker reuters detroit march 11
chrysler corp s 1 5 billion dlr bid to takeover
american motors corp should help bolster the
small automaker s sales but it leaves the future
of its 19 000 employees in doubt industry
analysts say it was business as usual yesterday
at the american ...
```

Index Construction

Stop words emphasized:

chrysler deal leaves uncertainty for amc workers by richard walker reuters detroit march 11 chrysler corp s 1 5 billion dlr bid to takeover american motors corp should help bolster the small automaker s sales but it leaves the future of its 19 000 employees in doubt industry analysts say it was business as usual yesterday at the american ...

Index Construction

After stemming:

```
chrysler deal leav uncertain amc work richard
walk reut detroit takeover american motor help
bols automak sal leav futur employ doubt industr
analy business usual yesterday american ...
```

Index Construction

After stemming:

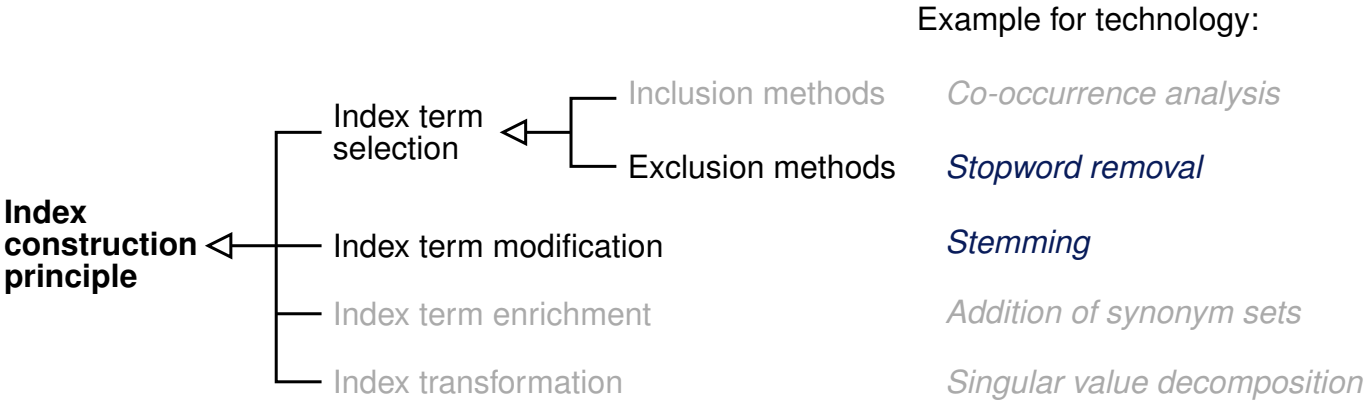
```
chrysler deal leav uncertain amc work richard
walk reut detroit takeover american motor help
bols automak sal leav futur employ doubt industr
analy business usual yesterday american ...
```

Vector Space Model:

$$\begin{pmatrix} \texttt{chrysler} \rightarrow 0.64 \\ \texttt{deal} \rightarrow 0.31 \\ \texttt{leav} \rightarrow 0.03 \\ \texttt{uncertain} \rightarrow 0.12 \\ \texttt{amc} \rightarrow 0.22 \\ \vdots \end{pmatrix}$$

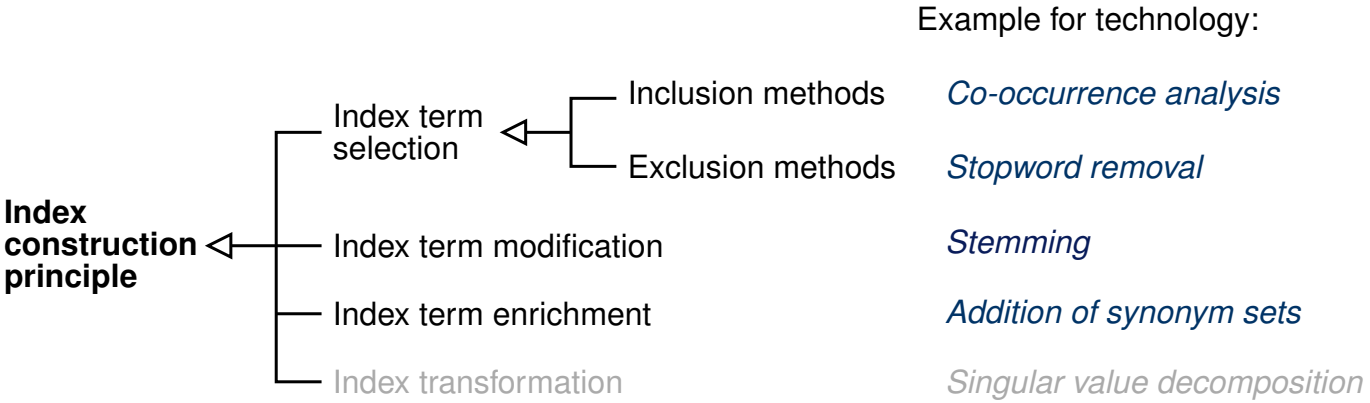
Term weighting schemes quantify the importance of each index term.

Index Construction Principles



How can the set of index terms be improved?

Enrichment Approaches



How can the set of index terms be improved?

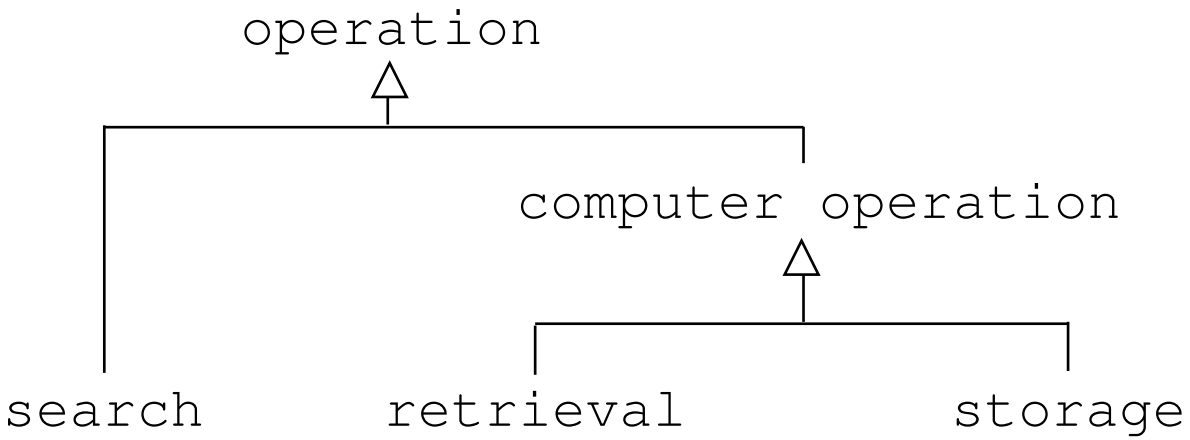
1. Semantic Approach.
   Exploit domain knowledge and external information sources to find or infer new index terms.

2. Syntactic Approach.
   Identify concepts (i.e. “Artificial Intelligence”) present in the document through statistical frequency analysis.

Enrichment Approaches

Semantic Approach: Find Transitive Relationships

Adding hypernyms:



Adding synonyms:

Synset for `message`:

{`content, subject matter, substance`}

[WordNet]

Enrichment Approaches

Syntactic Approach: Amplify Document Relationships

The area of information retrieval has grown well beyond its primary goals ...

... one of the most interesting and active areas of research in information retrieval.

... use common tools for the retrieval of parts or all of the deleted information.

Enrichment Approaches

Syntactic Approach: Amplify Document Relationships

`The area of information retrieval has grown well beyond its primary goals ...`

`... one of the most interesting and active areas of research in information retrieval.`

`... use common tools for the retrieval of parts or all of the deleted information.`

We consider a short sequence of words as a concept, if it has a particular meaning beyond the senses of each individual word.

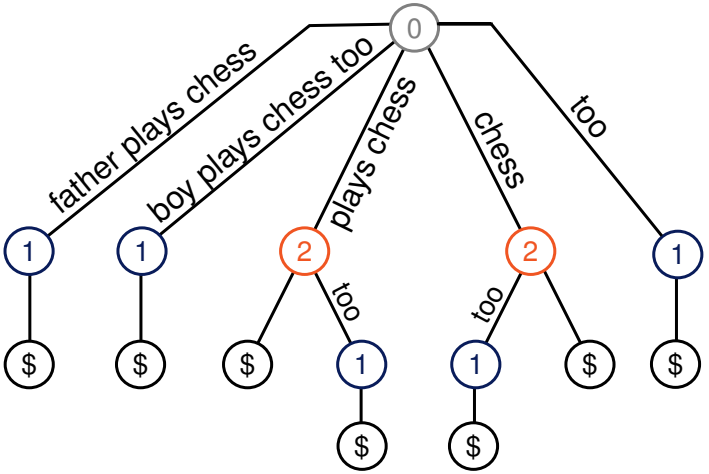
Concept identification:

Frequency analysis of all n-grams of a document, for $n \in \{2, 3, 4\}$.

Enrichment Approaches

Concept Identification: Successor Variety Analysis

Suffix tree at word level:



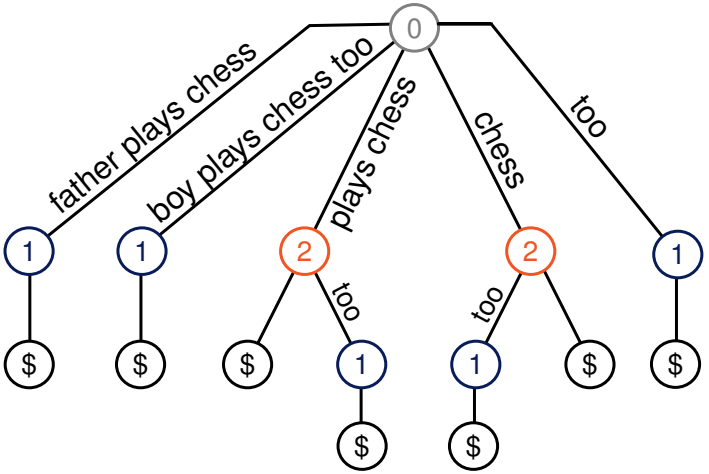
A note on runtime:

- $O(n)$ [Ukkonen 1995]
- $O(n^2)$ and $\Theta(n \log(n))$ [Giegerich et. al.]

Enrichment Approaches

Concept Identification: Successor Variety Analysis

Suffix tree at word level:



A note on runtime:

- $O(n)$ [Ukkonen 1995]
- $O(n^2)$ and $\Theta(n \log(n))$ [Giegerich et. al.]

How to find good candidates for a concept?

- analysis of degree differences (depending on tree depth)
- cut-off method, entropy method

Remark. Related work for stemming (suffix tree at letter level).

[Stein/Potthast 2006]

Enrichment Approaches

Concept Identification: Examples

Successor variety analysis at work:

$n = 2$	$n = 3$
south africa	mad cow disease
public sector	public sector deficit
european union	argentine central bank
weighted average	national statistics institute

$n = 4$
secretary general kofi annan
secretary state madeleine albright
prime minister benjamin netanyahu
palestinian president yasser arafat

Based on a sample of 1000 documents out of 5 categories from the RCV1.

Enrichment Approaches

Syntax vs. Semantics: Benefits and Weaknesses

Semantic Approach:

- \+ Transitive relationships are revealed
- – Generalization of specific documents
- – Word sense disambiguation may be necessary

Syntactic Approach:

- \+ Corpus-specific concepts are found
- \+ Language-independent means of concept identification
- – Statistical mass necessary to identify a concept

Evaluation

The Traditional Way: Clustering

Comparison of F-measure values:

Vector space model variant	F-min	F-max	F-av.
	(sample size 1000, 10 categories)		
standard vector space model	—baseline—		
synonym enrichment	-20%	+12%	-2%
hypernym enrichment	-9%	+20%	+3%
n-gram index term selection	0%	+14%	+8%

Evaluation

The Traditional Way: Clustering

Comparison of F-measure values:

Vector space model variant	F-min	F-max	F-av.
	(sample size 1000, 10 categories)		
standard vector space model	—baseline—		
synonym enrichment	-20%	+12%	-2%
hypernym enrichment	-9%	+20%	+3%
n-gram index term selection	0%	+14%	+8%

Interpretation is difficult.

A cluster algorithm's performance depends on various parameters.

Different cluster algorithms behave differently sensitive to document model "improvements".

Baseline? Interpretation? Objectivity? Generalizability?

Evaluation

Model-based instead of Algorithm-based: Expected Density $\bar{\rho}$

An objective way to rank document models is to compare their ability to *capture the intrinsic similarity relations* of a collection D.

Basic idea:

1. construct a similarity graph, $G = \langle V, E, w \rangle$
2. measure its conformance to a reference classification
3. analyze improvement/decline under new document model

Expected Density $\bar{\rho}$

Definition

Graph $G = \langle V, E, w \rangle$

- G is called sparse [dense] if $|E| = O(|V|)$ $[O(|V|^2)]$
- the density θ computes from the equation $|E| = |V|^\theta$

Expected Density $\bar{\rho}$

Definition

Graph $G = \langle V, E, w \rangle$

- G is called sparse [dense] if $|E| = O(|V|)$ $[O(|V|^2)]$
- the density θ computes from the equation $|E| = |V|^\theta$
- with $w(G) := \sum_{e \in E} w(e)$, this extends to weighted graphs:

$$w(G) = |V|^\theta \quad \Leftrightarrow \quad \theta = \frac{\ln(w(G))}{\ln(|V|)}$$

Using θ we assess the density of an induced subgraph G_i of G.

Expected Density $\bar{\rho}$

Definition

Graph $G = \langle V, E, w \rangle$

- G is called sparse [dense] if $|E| = O(|V|)$ $[O(|V|^2)]$
- the density θ computes from the equation $|E| = |V|^\theta$
- with $w(G) := \sum_{e \in E} w(e)$, this extends to weighted graphs:

$$w(G) = |V|^\theta \quad \Leftrightarrow \quad \theta = \frac{\ln(w(G))}{\ln(|V|)}$$

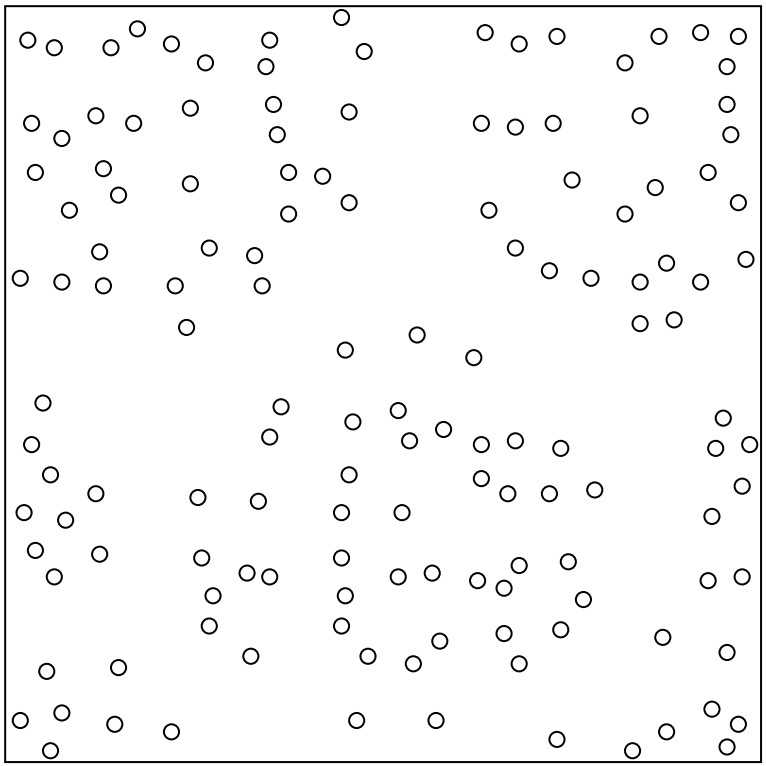
Using θ we assess the density of an induced subgraph G_i of G.

- a categorization $\mathcal{C} = \{C_1, \ldots, C_k\}$ induces k subgraphs G_i

➜ expected density $\overline{\rho}(\mathcal{C}) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^\theta}$

Expected Density $\bar{\rho}$

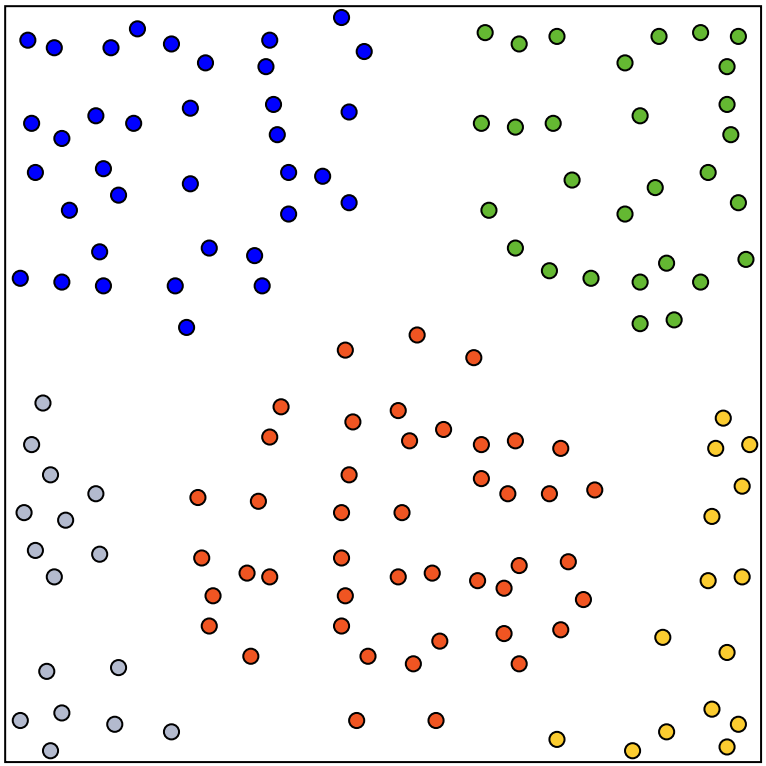
Understanding Expected Density



Embedding of a collection under a particular document model.

Expected Density $\bar{\rho}$

Understanding Expected Density

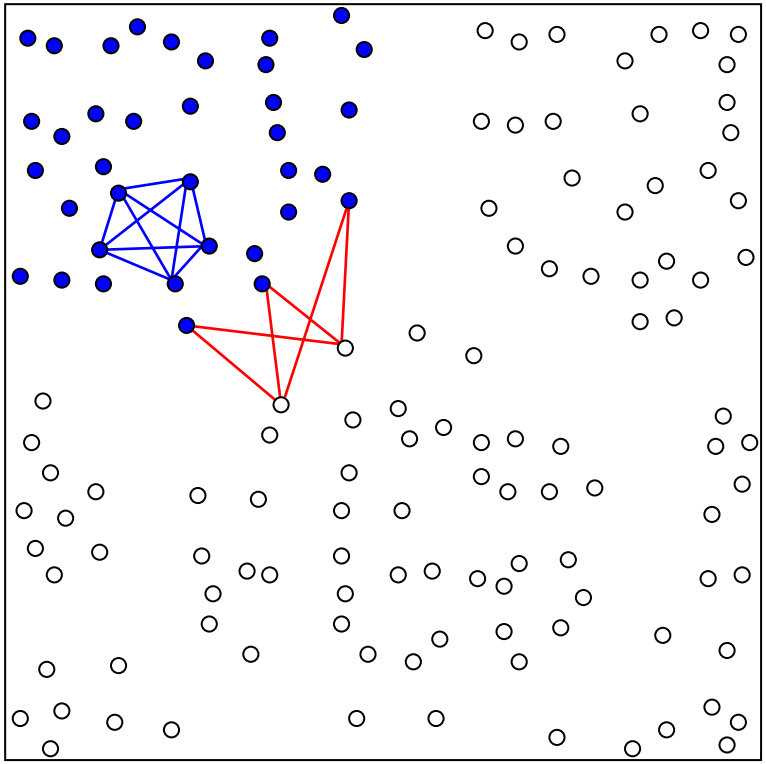


Embedding of a collection under a particular document model.

$\overline{\rho} > 1$ $[\overline{\rho} < 1]$ if the cluster density is larger [smaller] than average.

Expected Density $\bar{\rho}$

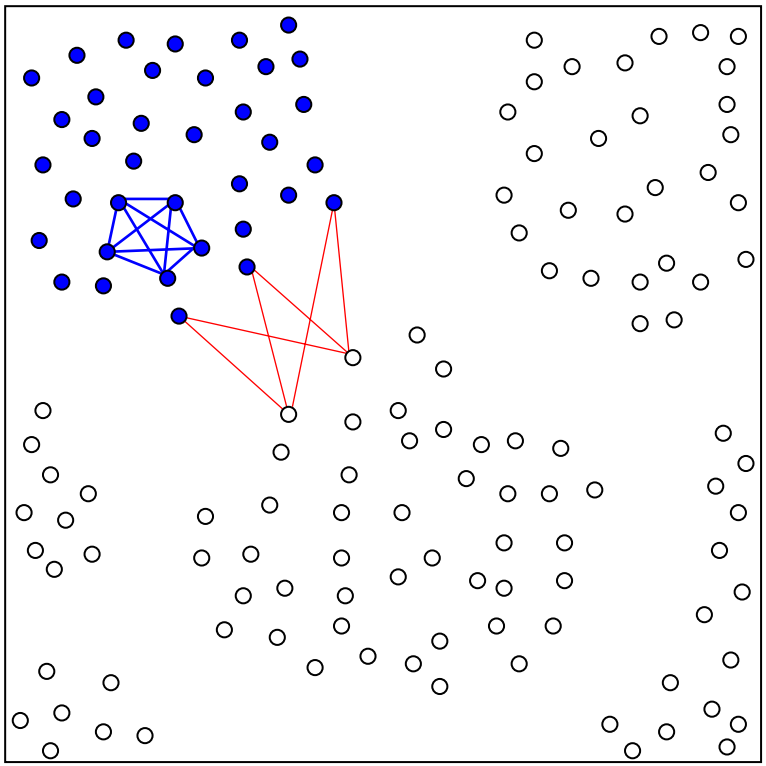
Understanding Expected Density



Consider inter-cluster and intra-cluster similarities.

Expected Density $\bar{\rho}$

Understanding Expected Density

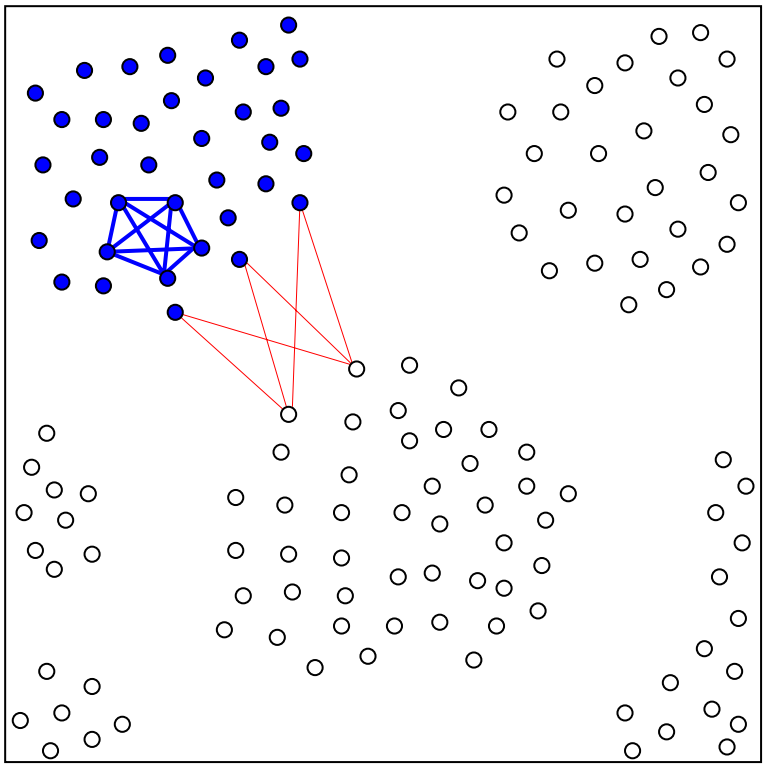


Consider inter-cluster and intra-cluster similarities.

Effect of a document model that *reinforces the structural characteristic* within a document collection.

Expected Density $\bar{\rho}$

Understanding Expected Density

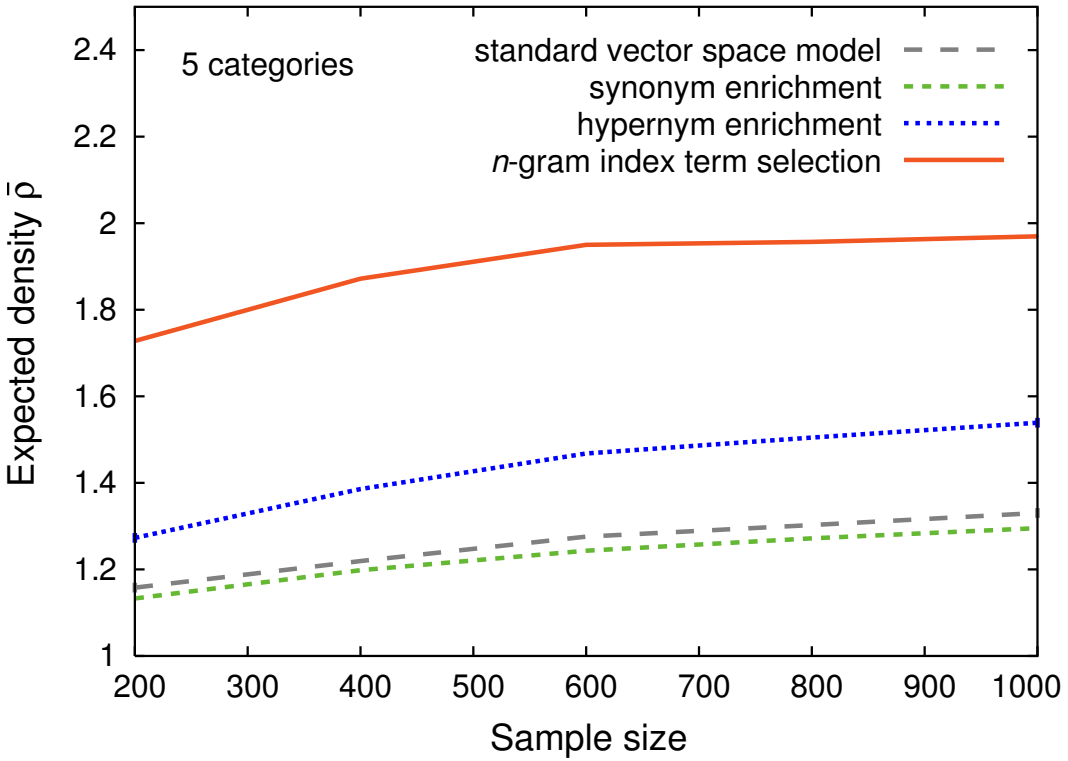


Consider inter-cluster and intra-cluster similarities.

Effect of a document model that *reinforces the structural characteristic* within a document collection.

Expected Density $\bar{\rho}$

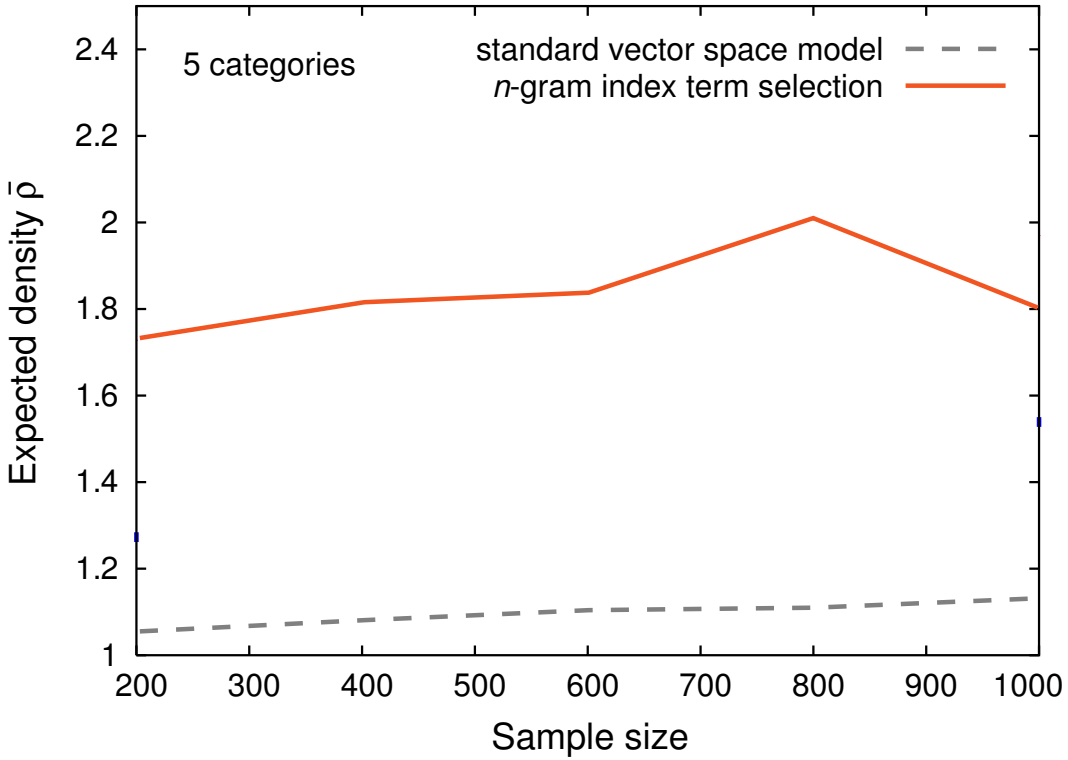
Experiments: English Collection



Collection: RCV1. Two documents d_1, d_2 are assigned to the same category if they share the top level category and the most specific category.

Expected Density $\bar{\rho}$

Experiments: German Collection



Collection: Compilation of 26,000 documents from 20 German news groups.

Summary

- Basis: document models with “visible” index terms
- Issue: selection, modification, enrichment of index terms
- Question: syntactic concept identification compared to semantic enrichment

Contribution

- efficient implementation of a concept identificator
- comparison to semantic enrichment apporaches
- algorithm-neutral evaluation method based on $\bar{\rho}$

Message

- the benefit of semantic term enrichment is overestimated
- generally accepted analysis methods are required