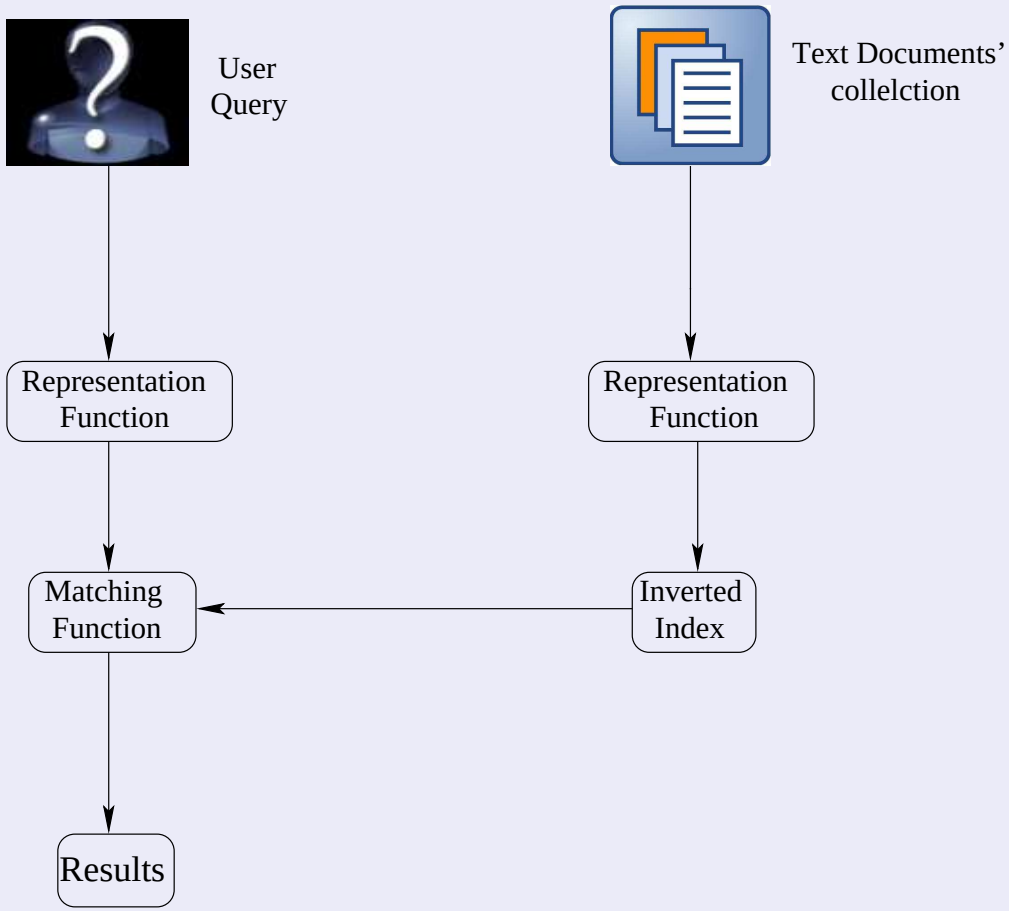
Analytic Knowledge Discovery Models for Information Retrieval and Text Summarization

Pawan Goyal

Team Sanskrit
INRIA Paris Rocquencourt

Outline

Ad-Hoc Information Retrieval



An Ad-Hoc Information Retrieval System

Information Retrieval: Datasets and Evaluation Criteria

TREC Dataset: Dataset used in Text REtrieval Conferences

- 741,686 documents, query topics 101-150 (TREC-2) and 151-200 (TREC-3).
- 524,000 documents, query topics 351-400 (TREC-7).
- Query example: *Topic 169: cost of garbage trash removal.*

Evaluation Criteria

- Precision at various points are computed. $P5 = \frac{N_{Rel}(5)}{5}$
- $N_{Rel}(x)$: Number of relevant documents in the top x documents returned by the system.
- Mean Averaged Precision (MAP) is the mean of the precision value at all recall points.
- Student's t-test is used to compare if the difference in results is statistically significant. (* : $p < 0.05$, ** : $p < 0.01$)

Ad-Hoc Information Retrieval

Problems that Ad-Hoc Information Retrieval addresses:

- Document and Query indexing: How to best represent their contents?

Ad-Hoc Information Retrieval

Problems that Ad-Hoc Information Retrieval addresses:

- Document and Query indexing: How to best represent their contents?
  $\rightarrow$ The representation allows real time search to be computationally efficient.

Ad-Hoc Information Retrieval

Problems that Ad-Hoc Information Retrieval addresses:

- Document and Query indexing: How to best represent their contents?
  $\rightarrow$ The representation allows real time search to be computationally efficient.
  $\rightarrow$ The representation minimizes the information loss.

Ad-Hoc Information Retrieval

Problems that Ad-Hoc Information Retrieval addresses:

- Document and Query indexing: How to best represent their contents?
  $\rightarrow$ The representation allows real time search to be computationally efficient.
  $\rightarrow$ The representation minimizes the information loss.
- Relevance measure: To what extent a document is relevant to a query?

Ad-Hoc Information Retrieval

Problems that Ad-Hoc Information Retrieval addresses:

- Document and Query indexing: How to best represent their contents?
  $\rightarrow$ The representation allows real time search to be computationally efficient.
  $\rightarrow$ The representation minimizes the information loss.
- Relevance measure: To what extent a document is relevant to a query?
- System evaluation: To what degree does the relevance measure reflect the human judgment?

Ad-Hoc Information Retrieval

Problems that Ad-Hoc Information Retrieval addresses:

- Document and Query indexing: How to best represent their contents?
  $\rightarrow$ The representation allows real time search to be computationally efficient.
  $\rightarrow$ The representation minimizes the information loss.
- Relevance measure: To what extent a document is relevant to a query?
- System evaluation: To what degree does the relevance measure reflect the human judgment?

Most Widely Used Approaches:

- Keyword based indexing to represent a document and a query
- Similarity measures such as Cosine similarity for relevance measures
- Precision and Recall measures for system evaluation

Essential factors for Keyword-based Indexing function F

- **Term frequency (f_{ij}):** How many times a term appear in a document?

Essential factors for Keyword-based Indexing function F

- **Term frequency (f_{ij}):** How many times a term appear in a document?

$$F \propto f_{ij}$$

Essential factors for Keyword-based Indexing function F

- **Term frequency (f_{ij}):** How many times a term appear in a document?

$$F \propto f_{ij}$$

- **Document length ($|D_i|$):** How many terms appear in the document?

Essential factors for Keyword-based Indexing function F

- **Term frequency (f_{ij}):** How many times a term appear in a document?

$$F \propto f_{ij}$$

- **Document length ($|D_i|$):** How many terms appear in the document?

$$F \propto \frac{1}{|D_i|}$$

Essential factors for Keyword-based Indexing function F

- **Term frequency (f_{ij}):** How many times a term appear in a document?

$$F \propto f_{ij}$$

- **Document length ($|D_i|$):** How many terms appear in the document?

$$F \propto \frac{1}{|D_i|}$$

- **Document Frequency (N_j):** Number of documents in which a term appears.

Essential factors for Keyword-based Indexing function F

- **Term frequency (f_{ij}):** How many times a term appear in a document?

$$F \propto f_{ij}$$

- **Document length ($|D_i|$):** How many terms appear in the document?

$$F \propto \frac{1}{|D_i|}$$

- **Document Frequency (N_j):** Number of documents in which a term appears.

$$F \propto \frac{1}{N_j}$$

Information Retrieval: An Example

Result of indexing

- Indexing: $D_1 \rightarrow \{(computation, 0.3), (information, 0.4), \ldots\}$, $D_2 \rightarrow \{(formal, 0.5), (computation, 0.4), \ldots\}, \ldots$

Information Retrieval: An Example

Result of indexing

- Indexing: $D_1 \rightarrow \{(computation, 0.3), (information, 0.4), \ldots\}$, $D_2 \rightarrow \{(formal, 0.5), (computation, 0.4), \ldots\}, \ldots$
- Inverted Index: $computation \rightarrow \{(D_1, 0.3), (D_2, 0.4), \ldots\}$

Information Retrieval: An Example

Result of indexing

- Indexing: $D_1 \rightarrow \{(computation, 0.3), (information, 0.4), \ldots\}$, $D_2 \rightarrow \{(formal, 0.5), (computation, 0.4), \ldots\}, \ldots$
- Inverted Index: $computation \rightarrow \{(D_1, 0.3), (D_2, 0.4), \ldots\}$
- User query, weighted by the system: $q \rightarrow \{(computation, 1), (information, 1), (processing, 1)\}$

Information Retrieval: An Example

Result of indexing

- Indexing: $D_1 \rightarrow \{(computation, 0.3), (information, 0.4), \ldots\}$, $D_2 \rightarrow \{(formal, 0.5), (computation, 0.4), \ldots\}, \ldots$
- Inverted Index: $computation \rightarrow \{(D_1, 0.3), (D_2, 0.4), \ldots\}$
- User query, weighted by the system: $q \rightarrow \{(computation, 1), (information, 1), (processing, 1)\}$
- $Sim(D_1, q) > Sim(D_2, q)$

Information Retrieval: An Example

Result of indexing

- Indexing: $D_1 \rightarrow \{(computation, 0.3), (information, 0.4), \ldots\}$, $D_2 \rightarrow \{(formal, 0.5), (computation, 0.4), \ldots\}, \ldots$
- Inverted Index: $computation \rightarrow \{(D_1, 0.3), (D_2, 0.4), \ldots\}$
- User query, weighted by the system: $q \rightarrow \{(computation, 1), (information, 1), (processing, 1)\}$
- $Sim(D_1, q) > Sim(D_2, q)$

Evaluation

- Let D_1 be relevant to q and D_2 be non-relevant as per human judgment.

Information Retrieval: An Example

Result of indexing

- Indexing: $D_1 \rightarrow \{(computation, 0.3), (information, 0.4), \ldots\}$, $D_2 \rightarrow \{(formal, 0.5), (computation, 0.4), \ldots\}, \ldots$
- Inverted Index: $computation \rightarrow \{(D_1, 0.3), (D_2, 0.4), \ldots\}$
- User query, weighted by the system: $q \rightarrow \{(computation, 1), (information, 1), (processing, 1)\}$
- $Sim(D_1, q) > Sim(D_2, q)$

Evaluation

- Let D_1 be relevant to q and D_2 be non-relevant as per human judgment.
- $P1 = 1.0$, $P2 = 0.5$, $MAP = 1.0$

Text Summarization

Why Text Summarization?

- Information Retrieval gives a list of documents, assumed to be relevant to the user query.
- There is still a vast volume of information in the retrieved documents.

Text Summarization

Why Text Summarization?

- Information Retrieval gives a list of documents, assumed to be relevant to the user query.
- There is still a vast volume of information in the retrieved documents.
- *Text summarization reduces this information into a short set of words or paragraph.*

Text Summarization

Why Text Summarization?

- Information Retrieval gives a list of documents, assumed to be relevant to the user query.
- There is still a vast volume of information in the retrieved documents.
- *Text summarization reduces this information into a short set of words or paragraph.*

Genres of Summary

- Extract vs. Abstract
  . . .lists fragments of text vs. re-phrases content coherently.
- Single document v/s Multi-document
  . . .based on one text vs. fuses together many texts.
- Generic v/s Query-oriented
  . . .provides author's view vs. reflects user's interest.

Text Summarization

Why Text Summarization?

- Information Retrieval gives a list of documents, assumed to be relevant to the user query.
- There is still a vast volume of information in the retrieved documents.
- *Text summarization reduces this information into a short set of words or paragraph.*

Genres of Summary

- Extract vs. Abstract
  . . .lists fragments of text vs. re-phrases content coherently.
- Single document v/s Multi-document
  . . .based on one text vs. fuses together many texts.
- Generic v/s Query-oriented
  . . .provides author's view vs. reflects user's interest.

Text Summarization

Generic Single document Extractive Summary

- Document is indexed along the same lines as for Information Retrieval.
- Similarity between sentences is represented using Cosine similarity.
- Important sentences are selected using PageRank based algorithm/eigen-values.

Text Summarization

Generic Single document Extractive Summary

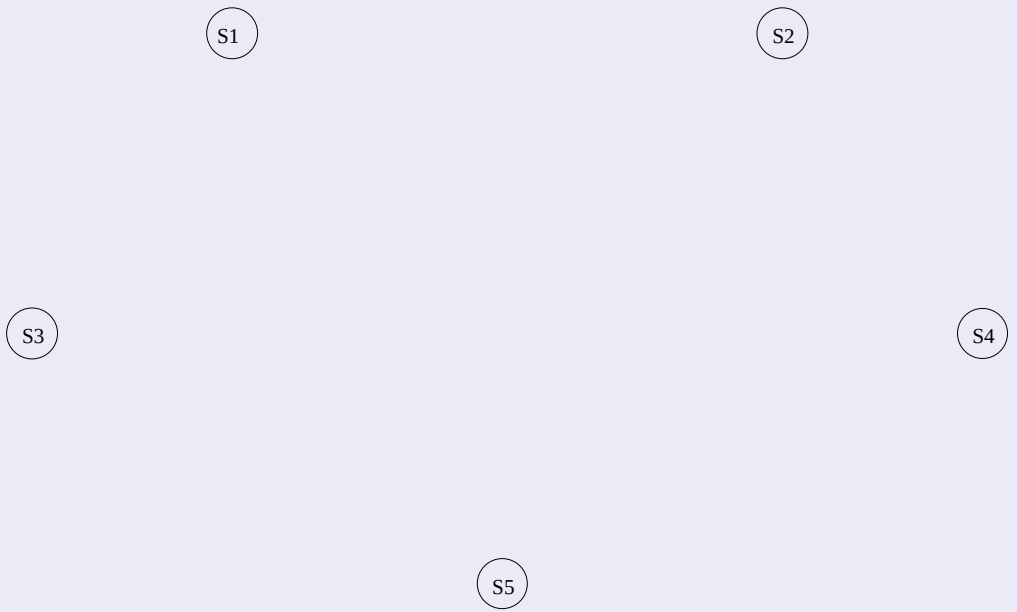
- Document is indexed along the same lines as for Information Retrieval.
- Similarity between sentences is represented using Cosine similarity.
- Important sentences are selected using PageRank based algorithm/eigen-values.

Evaluation Criteria

- DUC datasets: Various news articles used in Document Understanding Conferences.
- Manually created summaries are provided for each document.
- System generated summary is compared to the manually created summary.
- ROGUE toolkit is used for the evaluation.
- $$ROGUE - N = \frac{\sum_{S \in \{RefSum\}} \sum_{n-gram \in S} Count_{match}(n-gram)}{\sum_{S \in \{RefSum\}} \sum_{n-gram \in S} Count(n-gram)}$$

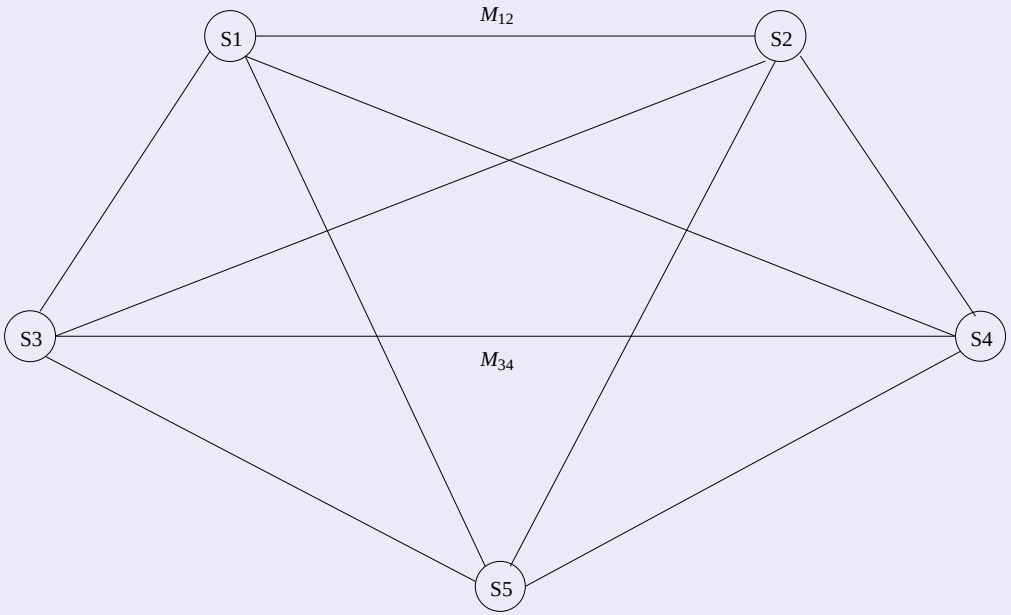
Text Summarization: Example

Sentence Graph



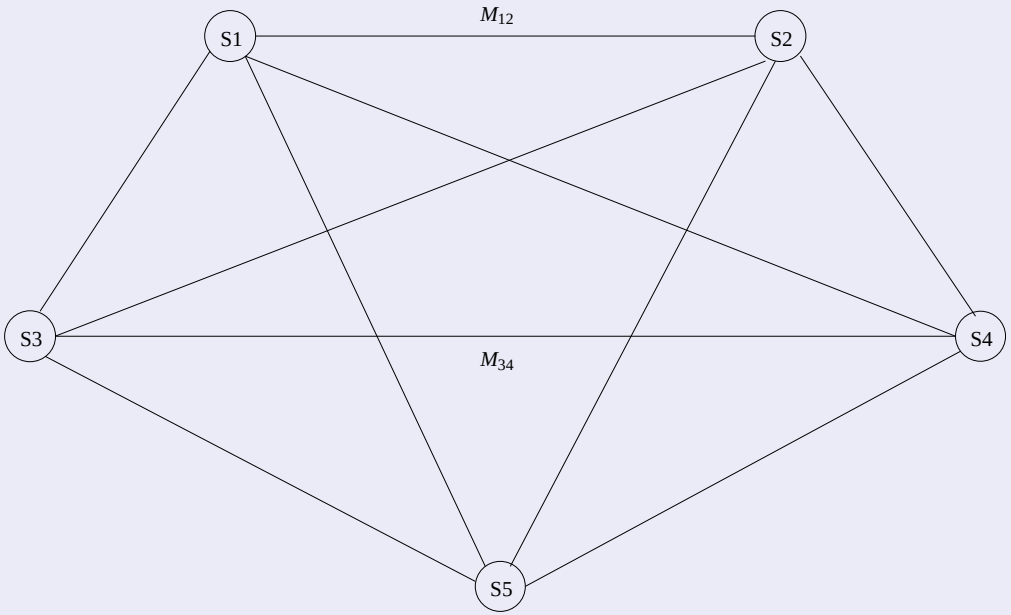
Text Summarization: Example

Sentence Graph



Text Summarization: Example

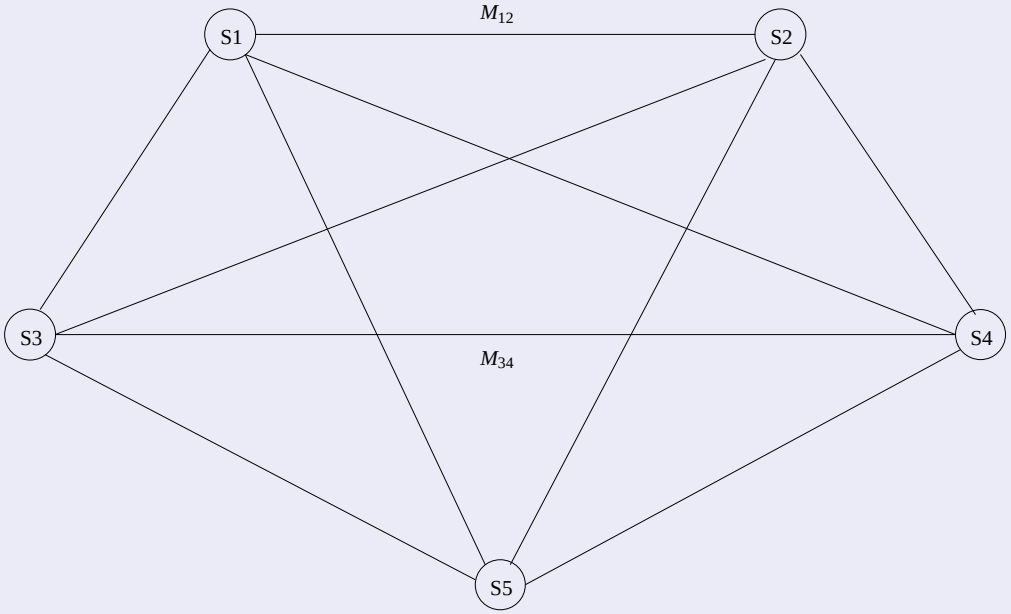
Sentence Graph



$$\tilde{M} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.4 & 0.1 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.4 & 0.0 & 0.4 & 0.0 & 0.2 \\ 0.3 & 0.0 & 0.0 & 0.7 & 0.0 \end{bmatrix}$$

Text Summarization: Example

Sentence Graph



$$\tilde{M} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.4 & 0.1 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.4 & 0.0 & 0.4 & 0.0 & 0.2 \\ 0.3 & 0.0 & 0.0 & 0.7 & 0.0 \end{bmatrix}$$

Solving using Page-Rank based algorithm iteratively for sentence centrality vector I:

$$I_j = \mu \cdot \sum_{\forall k \neq j} I_k \cdot \tilde{M}_{j,k} + \frac{1-\mu}{|S|}$$

$$I = \begin{bmatrix} 0.22 & 0.18 & 0.2 & 0.3 & 0.1 \end{bmatrix}$$

Research Problems

Term Mismatch

- Stems from the word independence assumption
- User query: *insurance cover which pays for long term care.*
- A relevant document may contain terms different from the actual user query.
- Some relevant words concerning this query: $\{medicare, premiums, insurers\}$

Research Problems

Term Mismatch

- Stems from the word independence assumption
- User query: *insurance cover which pays for long term care.*
- A relevant document may contain terms different from the actual user query.
- Some relevant words concerning this query: $\{medicare, premiums, insurers\}$

Existing Solutions

- Manually constructed ontologies such as Wordnet
- Relevance feedback
- Co-occurrence models such as mutual information

Research Problems

Context Independent Word Indexing

Information Retrieval

$$D_1 = \{\text{robot}, \text{healthcare}, \text{mobile}, \text{autonomous}, \text{research}\}$$
$$D_2 = \{\text{fifa}, \text{soccer}, \text{germany}, \text{played}, \text{robot}\}$$

Research Problems

Context Independent Word Indexing

Information Retrieval

$$D_1 = \{\text{robot, healthcare, mobile, autonomous, research}\}$$
$$D_2 = \{\text{fifa, soccer, germany, played, robot}\}$$

Sentence Extraction

$$D_1 : S_{11} = \{\text{started, career, engineering}\}$$
$$: S_{12} = \{\text{shifted, engineering, humanities}\}$$
$$D_2 : S_{21} = \{\text{engineering, application, scientific, principles}\}$$
$$: S_{22} = \{\text{engineering, design, build, machines}\}$$

Research Problems

Context Independent Word Indexing

Information Retrieval

$$D_1 = \{\text{robot, healthcare, mobile, autonomous, research}\}$$
$$D_2 = \{\text{fifa, soccer, germany, played, robot}\}$$

Sentence Extraction

$$D_1 : S_{11} = \{\text{started, career, engineering}\}$$
$$: S_{12} = \{\text{shifted, engineering, humanities}\}$$
$$D_2 : S_{21} = \{\text{engineering, application, scientific, principles}\}$$
$$: S_{22} = \{\text{engineering, design, build, machines}\}$$

Existing Solutions

- Document Clustering
- Latent Semantic Analysis

Knowledge Discovery: A Potential solution

Knowledge Discovery: A Potential solution

Distributional Hypothesis

"You know a word by the company it keeps." (Firth, 1957)

Knowledge Discovery: A Potential solution

Distributional Hypothesis

"You know a word by the company it keeps." (Firth, 1957)

"Words that occur in the same contexts tend to have similar meanings." (Zellig Harris, 1968)

Knowledge Discovery: A Potential solution

Distributional Hypothesis

"You know a word by the company it keeps." (Firth, 1957)

"Words that occur in the same contexts tend to have similar meanings." (Zellig Harris, 1968)

$\rightarrow$ Semantically similar words tend to have similar distributional patterns.

Knowledge Discovery: A Potential solution

Distributional Hypothesis

"You know a word by the company it keeps." (Firth, 1957)

"Words that occur in the same contexts tend to have similar meanings." (Zellig Harris, 1968)

$\rightarrow$ Semantically similar words tend to have similar distributional patterns.

My Approach: Specific Objectives

- Using distributional hypothesis to analyze the research problems from a theoretical perspective.
- To empirically evaluate the proposed analytic knowledge discovery models with respect to the existing approaches.

Query Representation Model: Basic Idea

Distributional Hypothesis

Words are not independent of each other

'computation' provides more information about 'algorithm' and 'programming' than about 'petroleum' or 'environment'.

Query Representation Model: Basic Idea

Distributional Hypothesis

Words are not independent of each other

'computation' provides more information about 'algorithm' and 'programming' than about 'petroleum' or 'environment'.

$\rightarrow$ *Distributional pattern of terms is used to find the terms, related to the query words.*

Query Representation Model: Basic Idea

Distributional Hypothesis

Words are not independent of each other

'computation' provides more information about 'algorithm' and 'programming' than about 'petroleum' or 'environment'.

$\rightarrow$ *Distributional pattern of terms is used to find the terms, related to the query words.*

Compositional Model

Problem of 'polysemy'

'mouse' can provide information about $\{keyboard, monitor\}$ in one context and about $\{animals, food\}$ in other context.

Query Representation Model: Basic Idea

Distributional Hypothesis

Words are not independent of each other

'computation' provides more information about 'algorithm' and 'programming' than about 'petroleum' or 'environment'.

$\rightarrow$ *Distributional pattern of terms is used to find the terms, related to the query words.*

Compositional Model

Problem of 'polysemy'

'mouse' can provide information about $\{keyboard, monitor\}$ in one context and about $\{animals, food\}$ in other context.

$\rightarrow$ *Combined effect of all query terms is used to avoid 'polysemy': $\{mouse, wireless\}$ can disambiguate the two usages of mouse.*

Query Representation Model for 'Term Mismatch' problem

The parametric model

- A: Matrix that captures the distributional pattern
- Let a user query be represented as $Q = q.A$

Query Representation Model for 'Term Mismatch' problem

The parametric model

- A: Matrix that captures the distributional pattern
- Let a user query be represented as $Q = q.A$

  $\rightarrow$ *What choice of* A_{ij} *will give an enhanced retrieval performance?*

Query Representation Model for 'Term Mismatch' problem

The parametric model

- A: Matrix that captures the distributional pattern
- Let a user query be represented as $Q = q.A$

  $\rightarrow$ *What choice of* A_{ij} *will give an enhanced retrieval performance?*
- $\boxed{A_{ij} = f\left(\frac{\sum_i \sum_j t_{ij} \cdot \sum_k (\delta_{ki} t_{kj})}{\sum_k t_{ki} \cdot \sum_k t_{kj}}\right)}$: Derived using system relevance criteria and an empirical evidence from user relevance criteria.
- $\boxed{f = log(x + 0.05)}$: Fixed through sensitivity analysis.

Query Representation Model for 'Term Mismatch' problem

The parametric model

- A: Matrix that captures the distributional pattern
- Let a user query be represented as $Q = q.A$

  $\rightarrow$ *What choice of* A_{ij} *will give an enhanced retrieval performance?*
- $\boxed{A_{ij} = f\left(\frac{\sum_i \sum_j t_{ij} \cdot \sum_k (\delta_{ki} t_{kj})}{\sum_k t_{ki} \cdot \sum_k t_{kj}}\right)}$: Derived using system relevance criteria and an empirical evidence from user relevance criteria.
- $\boxed{f = log(x + 0.05)}$: Fixed through sensitivity analysis.

The only query expansion model with a relevance based justification.

Results

TREC Topic 104: catastrophic health insurance

Query Representation: surtax:1.0 hcfa:0.97 medicare:0.93 hmos:0.83 medicaid:0.8 hmo:0.78 beneficiaries:0.75 ambulatory:0.72 premiums:0.72 hospitalization:0.71 hhs:0.7 reimbursable:0.7 deductible:0.69

Results

TREC Topic 104: catastrophic health insurance

Query Representation: surtax:1.0 hcfa:0.97 medicare:0.93 hmos:0.83 medicaid:0.8 hmo:0.78 beneficiaries:0.75 ambulatory:0.72 premiums:0.72 hospitalization:0.71 hhs:0.7 reimbursable:0.7 deductible:0.69

- Broad expansion terms: **medicare, beneficiaries, premiums** . . .
- Specific domain terms: **HCFA** (Health Care Financing Administration), **HMO** (Health Maintenance Organization), **HHS** (Health and Human Services)

Results

TREC Topic 104: catastrophic health insurance

Query Representation: surtax:1.0 hcfa:0.97 medicare:0.93 hmos:0.83 medicaid:0.8 hmo:0.78 beneficiaries:0.75 ambulatory:0.72 premiums:0.72 hospitalization:0.71 hhs:0.7 reimbursable:0.7 deductible:0.69

- Broad expansion terms: **medicare, beneficiaries, premiums** . . .
- Specific domain terms: **HCFA** (Health Care Financing Administration), **HMO** (Health Maintenance Organization), **HHS** (Health and Human Services)

TREC Topic 355: ocean remote sensing

Query Representation: radiometer:1.0 landsat:0.97 ionosphere:0.94 cnes:0.84 altimeter:0.83 nasda:0.81 meterology:0.81 cartography:0.78 geostationary:0.78 doppler:0.78 oceanographic:0.76

Results

TREC Topic 104: catastrophic health insurance

Query Representation: surtax:1.0 hcfa:0.97 medicare:0.93 hmos:0.83 medicaid:0.8 hmo:0.78 beneficiaries:0.75 ambulatory:0.72 premiums:0.72 hospitalization:0.71 hhs:0.7 reimbursable:0.7 deductible:0.69

- Broad expansion terms: **medicare, beneficiaries, premiums** ...
- Specific domain terms: **HCFA** (Health Care Financing Administration), **HMO** (Health Maintenance Organization), **HHS** (Health and Human Services)

TREC Topic 355: ocean remote sensing

Query Representation: radiometer:1.0 landsat:0.97 ionosphere:0.94 cnes:0.84 altimeter:0.83 nasda:0.81 meterology:0.81 cartography:0.78 geostationary:0.78 doppler:0.78 oceanographic:0.76

- Broad expansion terms: **radiometer, landsat, ionosphere** ...
- Specific domain terms: **CNES** (Centre National dÉtudes Spatiales) and **NASDA** (National Space Development Agency of Japan)

Neighborhood Based Document Smoothing (NBDS) Model

Context Sensitive Document Indexing

$D_1 = \{$**robot**, healthcare, mobile, autonomous, research$\}$

$D_2 = \{$fifa, soccer, germany, played, **robot**$\}$

- Content-carrying (Topical) terms should be given higher weights than the background terms.
- Topical terms are supposed to have higher association with each other, when computed on a large corpora.
- $\boxed{{t_{ij}}^N = \beta t_{ij} + \gamma \sum_k (A_{jk} t_{ik})}$: Proposed model to redistribute the indexing weights.

Neighborhood Based Document Smoothing (NBDS) Model

Context Sensitive Document Indexing

$D_1 = \{\textbf{robot}, \text{healthcare, mobile, autonomous, research}\}$

$D_2 = \{\text{fifa, soccer, germany, played, } \textbf{robot}\}$

- Content-carrying (Topical) terms should be given higher weights than the background terms.
- Topical terms are supposed to have higher association with each other, when computed on a large corpora.
- $\boxed{{t_{ij}}^N = \beta t_{ij} + \gamma \sum_k (A_{jk} t_{ik})}$: Proposed model to redistribute the indexing weights.

NBDS Model: Main Features

- The model does not cause any extra computational burden at run-time.
- The only model which provides a mathematical framework with a relevance-based justification.

A Context Based Word Indexing Model for Text summarization

Bernoulli model of co-occurrence for lexical association

- Consider the distribution of terms t_i and t_j in a corpus of N documents.
- N_i, N_j: Number of documents in which t_i and t_j occur respectively.
- N_{ij}: Number of documents in which t_i and t_j co-occur.
- Probability p_i of the term t_i appearing in an arbitrary document: $\boxed{p_i = \frac{N_i}{N}}$
- Term t_i occurs in N_{ij} documents out of these N_j documents and does not occur in $N_j - N_{ij}$ documents.
- Using Bernoulli distribution: $\boxed{p(N_{ij}) = \binom{N_j}{N_{ij}} {p_i}^{N_{ij}} {q_i}^{N_j - N_{ij}}}$
- Using Shannon's self-information notion: $\boxed{Inf(N_{ij}) = -log_2(p(N_{ij}))}$
- Stirling's approximation: $\boxed{n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n}$
- $Inf(N_{ij})$ is used to modify the indexing weights iteratively.

Results

Comparison of Query Representation over the Language Model

Dataset		LM	CQE	MCTM	QR (Improvements %)
TREC-2	MAP	0.183	0.192	0.185	0.203 (+10.9**,+5.7,+9.7*)
	P30	0.386	0.393	0.392	0.415 (+7.5,+5.6,+5.9)
TREC-7	MAP	0.179	0.184	0.184	0.2 (+11.7**,+8.7**,+8.7*)
	P30	0.289	0.284	0.291	0.315 (+9.0**,+10.9*,+8.2*)

Comparison of NBDS Model applied to the Language model

Dataset		LM	LM+NBDS	Improvement (%)
TREC 2	MAP	0.183	0.199	+8.7**
	P10	0.448	0.462	+3.1
TREC 3	MAP	0.197	0.212	+7.6**
	P10	0.474	0.53	+11.8**

Results

Sentence Extraction Experiments

System	DUC01		DUC02	
	ROGUE-1	ROGUE-2	ROGUE-1	ROGUE-2
IntraLink	0.439	0.172	0.45	0.19
IntraLink+bern	0.447	0.184	0.461	0.202
UniformLink	0.438	0.173	0.458	0.199
UniformLink+bern	0.443	0.183	0.462	0.205

Conclusions

- The problems of ‘term mismatch’ and ‘context independent document indexing’ have been addressed using distributional hypothesis.
- A proper mathematical framework has been provided to the query expansion and document smoothing techniques.
- The proposed knowledge discovery models have been shown to perform significantly superior to the traditional retrieval frameworks.
- Being developed in the generalized retrieval framework, these models are applicable to all of the retrieval frameworks.
- The proposed models for document smoothing do not cause any extra computational burden at run-time.

Publications

- Pawan Goyal, Laxmidhar Behera and T. M. McGinnity, "*Query Representation through Lexical Association for Information Retrieval*" (Preprint) IEEE Transactions on Knowledge and Data Engineering, vol. PP, July 2011.
- Pawan Goyal, Laxmidhar Behera and T. M. McGinnity, "*A Novel Neighborhood Based Document Smoothing Model for Information Retrieval*", Revised and Submitted to Information Retrieval, Springer.
- Pawan Goyal, Laxmidhar Behera and T. M. McGinnity, "*A Context Based Word Indexing Model for Text Summarization*" Submitted to IEEE Transactions on Knowledge and Data Engineering.
- Pawan Goyal, Laxmidhar Behera and T. M. McGinnity, "*Application of Bayesian Framework in Natural Language Understanding*", IETE Technical Review, Volume 25, Issue 5, pp 251-269, 2008 .
- Pawan Goyal, Laxmidhar Behera and T. M. McGinnity, "*Entailment of Causal Queries in Narratives Using Action Language*". In the proceedings of KDIR 2009, October 04-06, Funchal, Portugal, pp 112-118.
- Pawan Goyal, Laxmidhar Behera and T. M. McGinnity, "*An Information Retrieval Model Based On Automatically Learnt Concept Hierarchies*". In the proceedings of IEEE ICSC 2009, Berkeley, CA, USA, pp 458-465.
- Pawan Goyal, Laxmidhar Behera and T. M. McGinnity, "*An Information Retrieval Approach Based on Semantically Adapted Vector Space Model*". In the proceedings of ICON, 2009, December 14-17, Hyderabad, INDIA.

Questions?