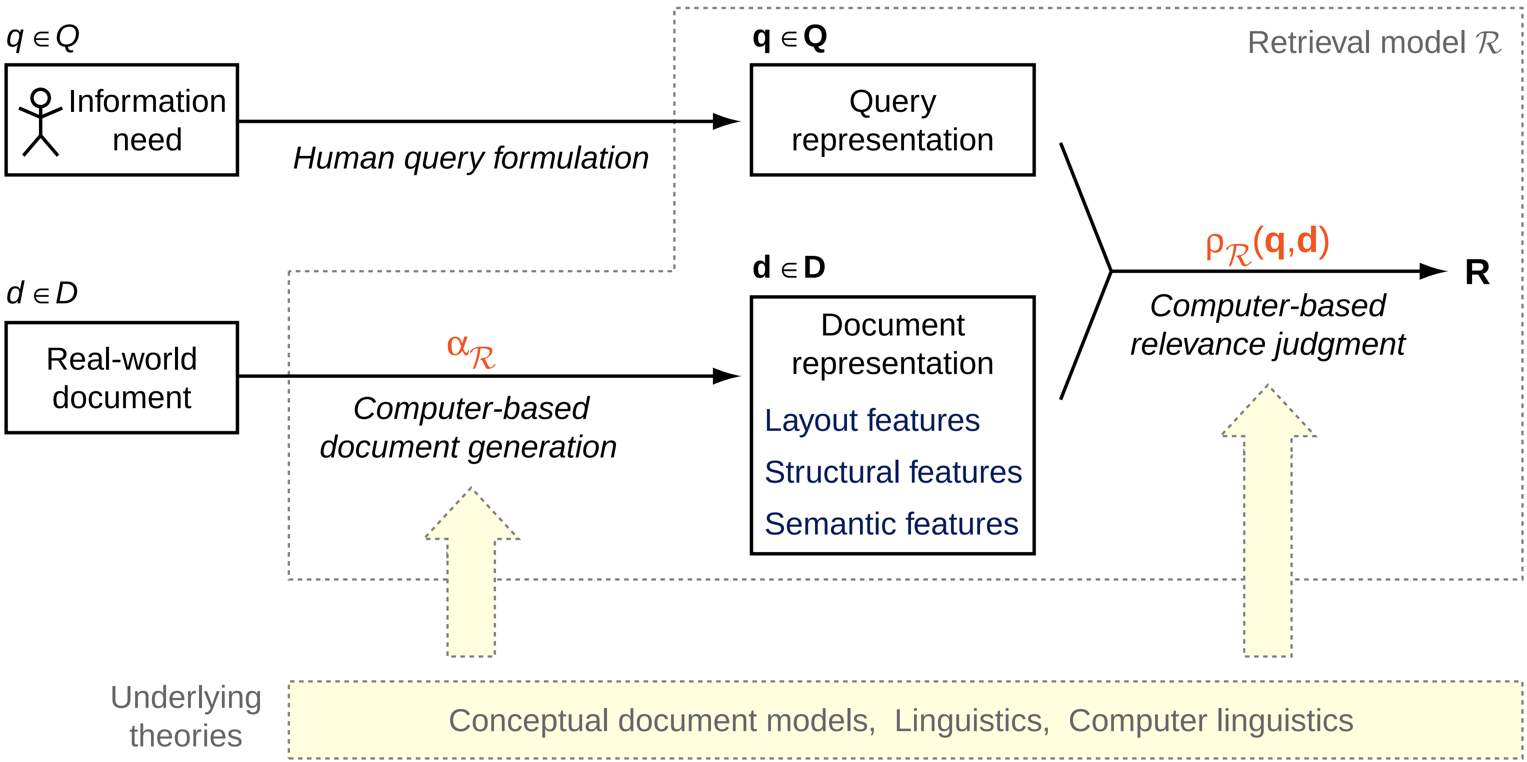
Collection-Relative Representations

A Unifying View to Retrieval Models

Benno Stein Maik Anderka
Bauhaus-Universität Weimar
www.webis.de

Retrieval Models



Outline

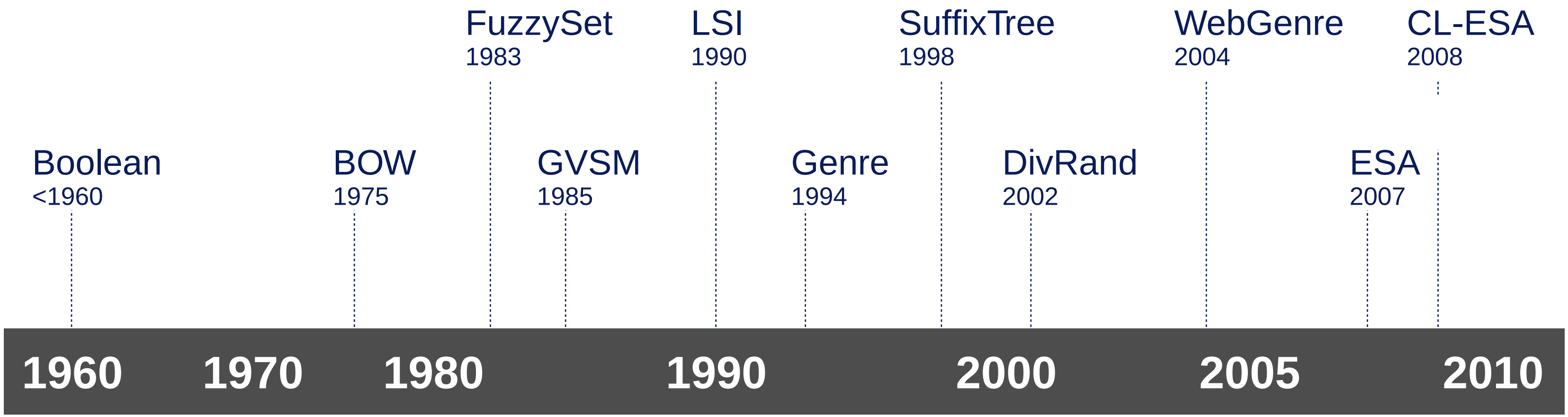
- Retrieval Model Timeline
- The ESA Model Revisited
- Framework of Collection-Relative Retrieval Models

Retrieval Models

Retrieval Models

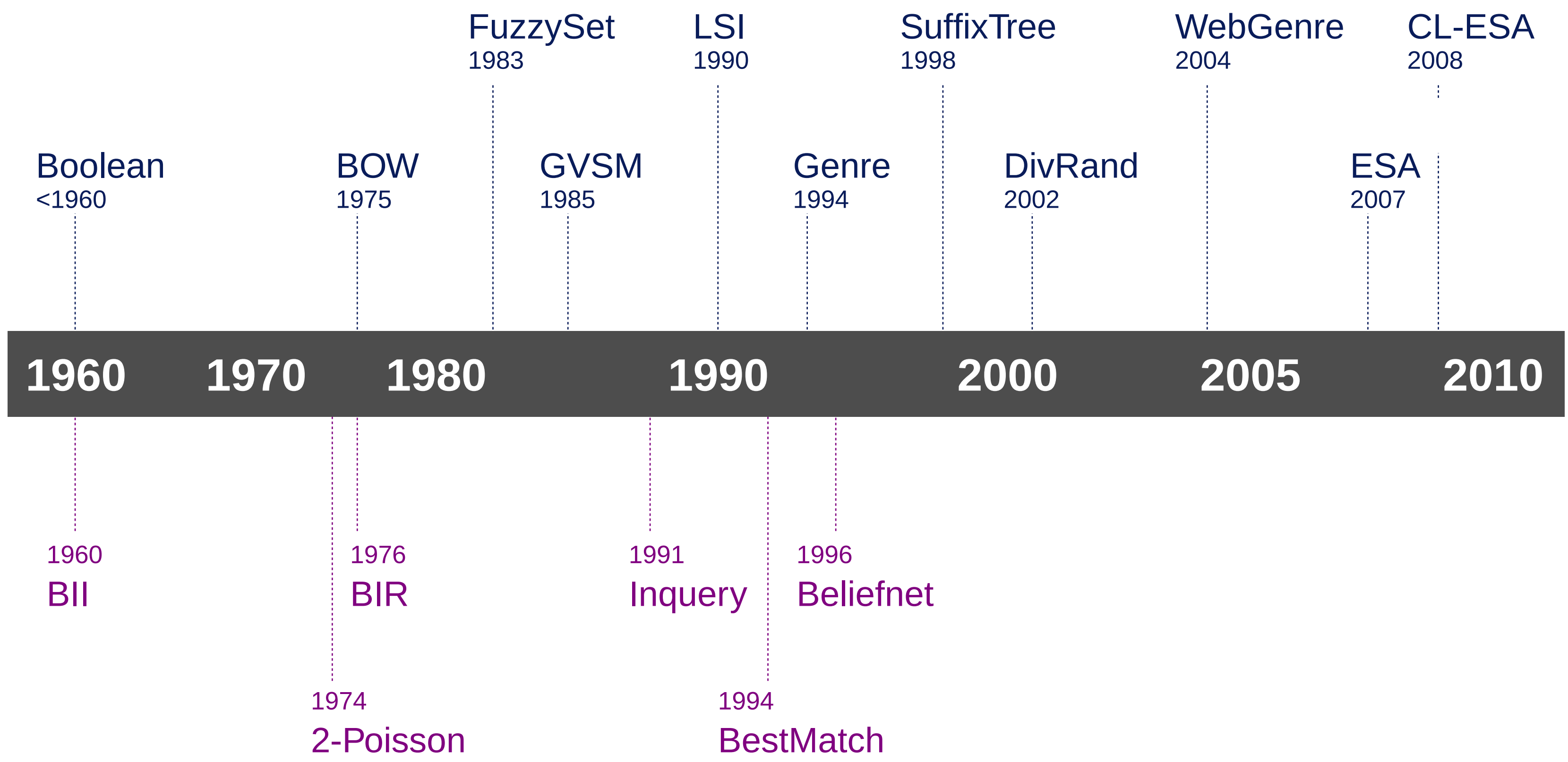
1960	1970	1980	1990	2000	2005	2010

Retrieval Models



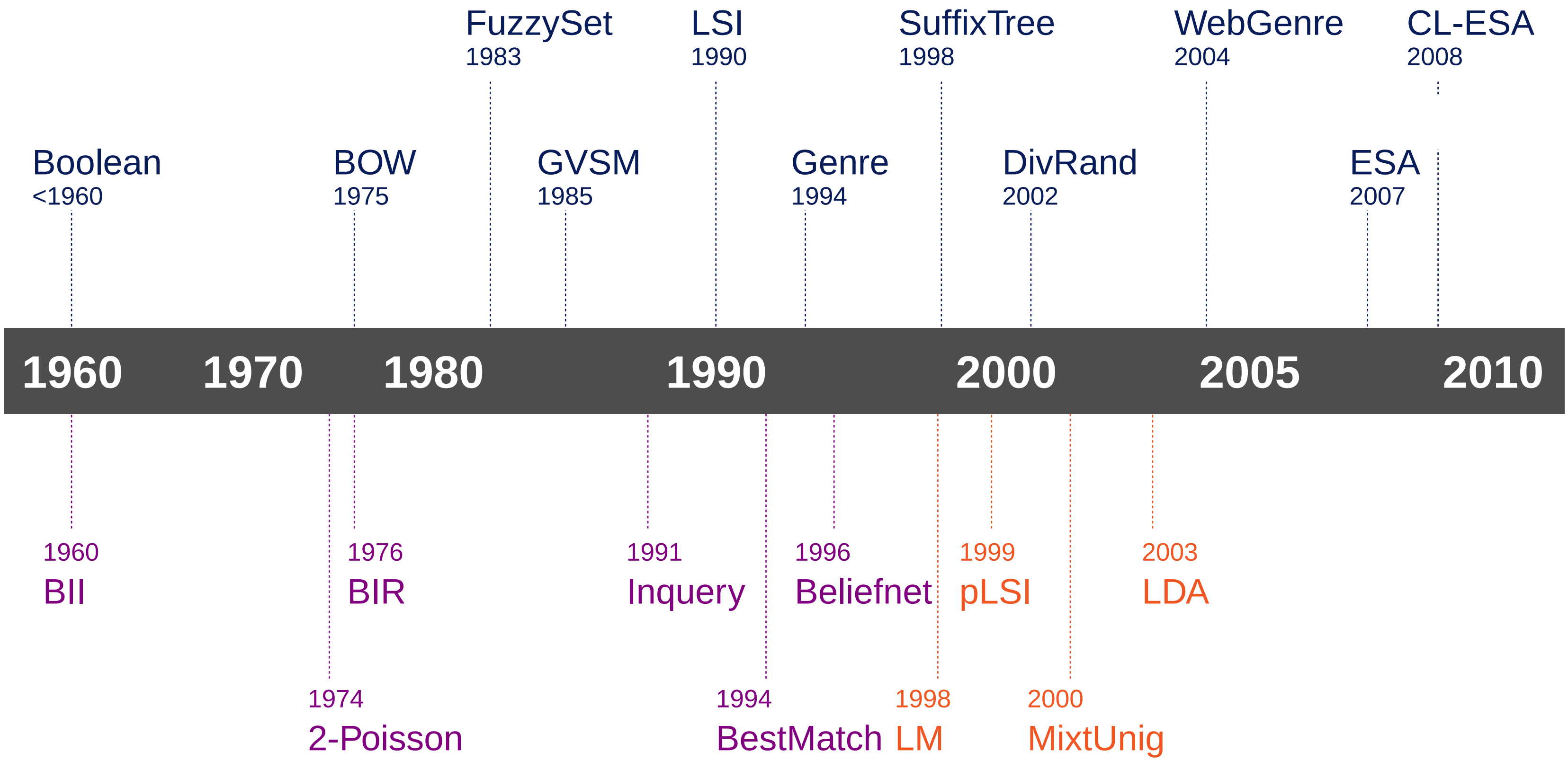
- Empirical models

Retrieval Models



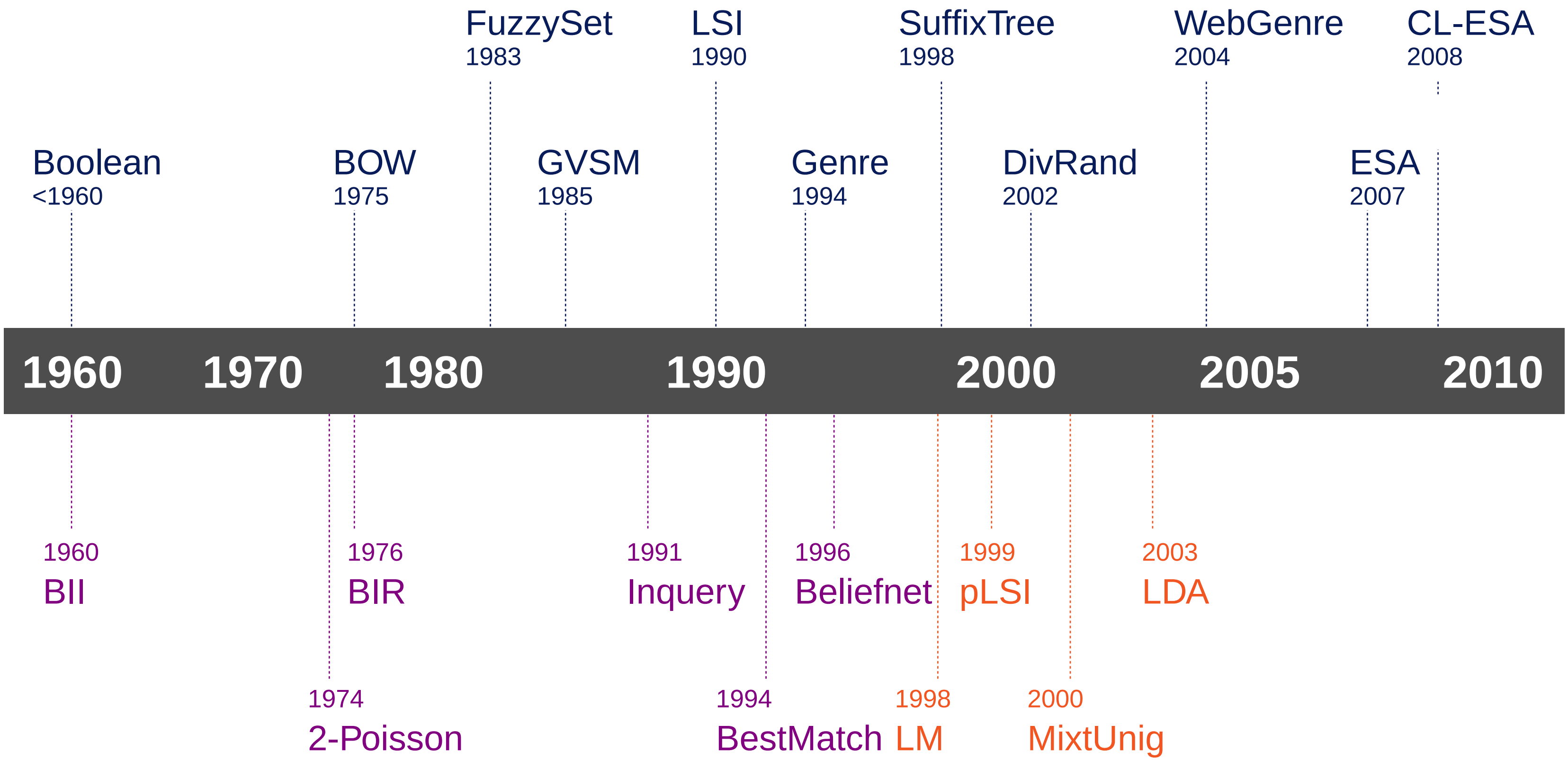
- Empirical models
- Probabilistic models

Retrieval Models



- Empirical models
- Probabilistic models
- Language models

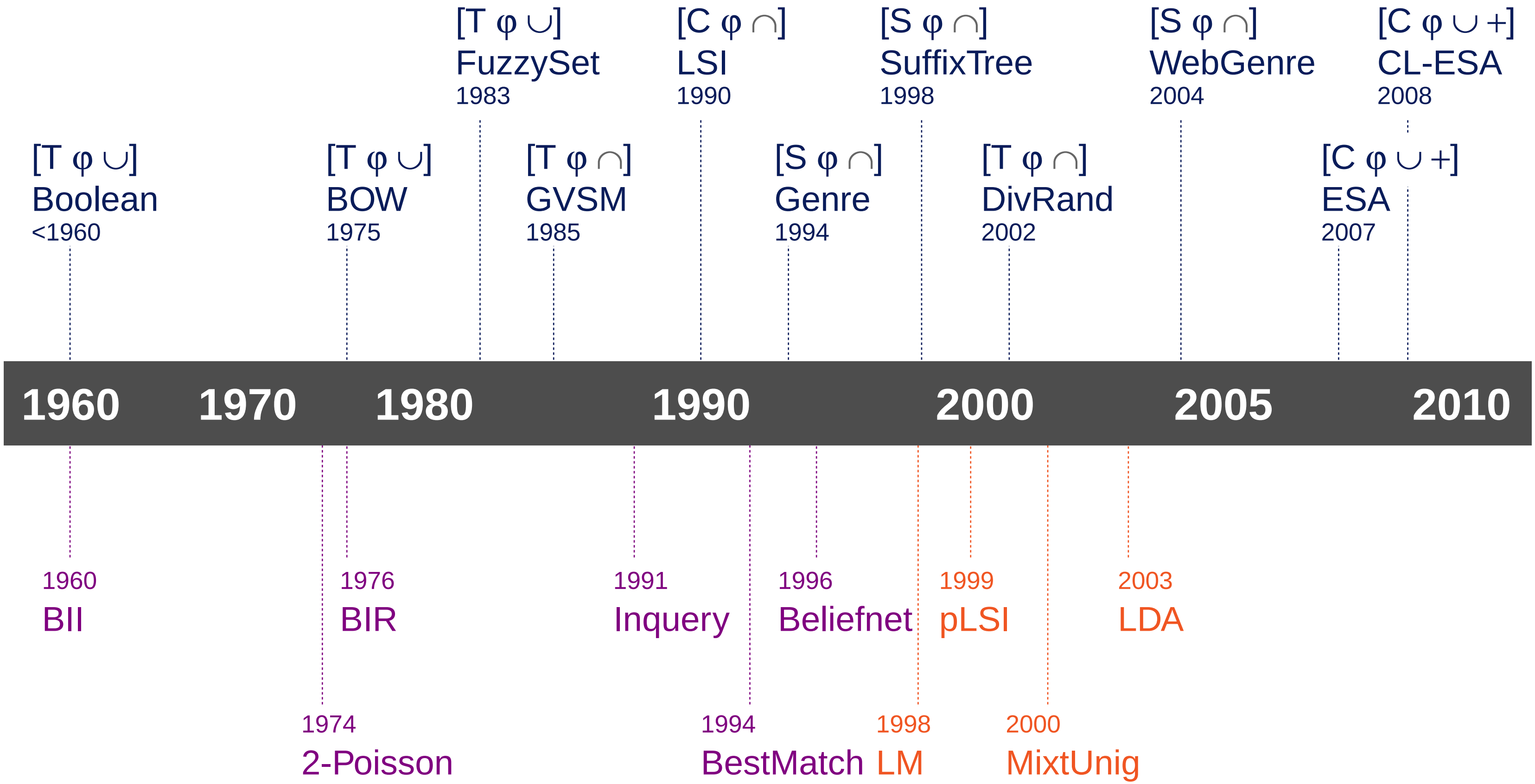
Retrieval Models



- Empirical models
- Probabilistic models
- Language models

Feature space	[T] terms [C] concepts [S] special
RSV foundation	[φ] sim. [ρ] relevance [γ] generation
Collection	[∪] open [∩] closed
Ext. knowledge	[✓] user feedback [+] collection

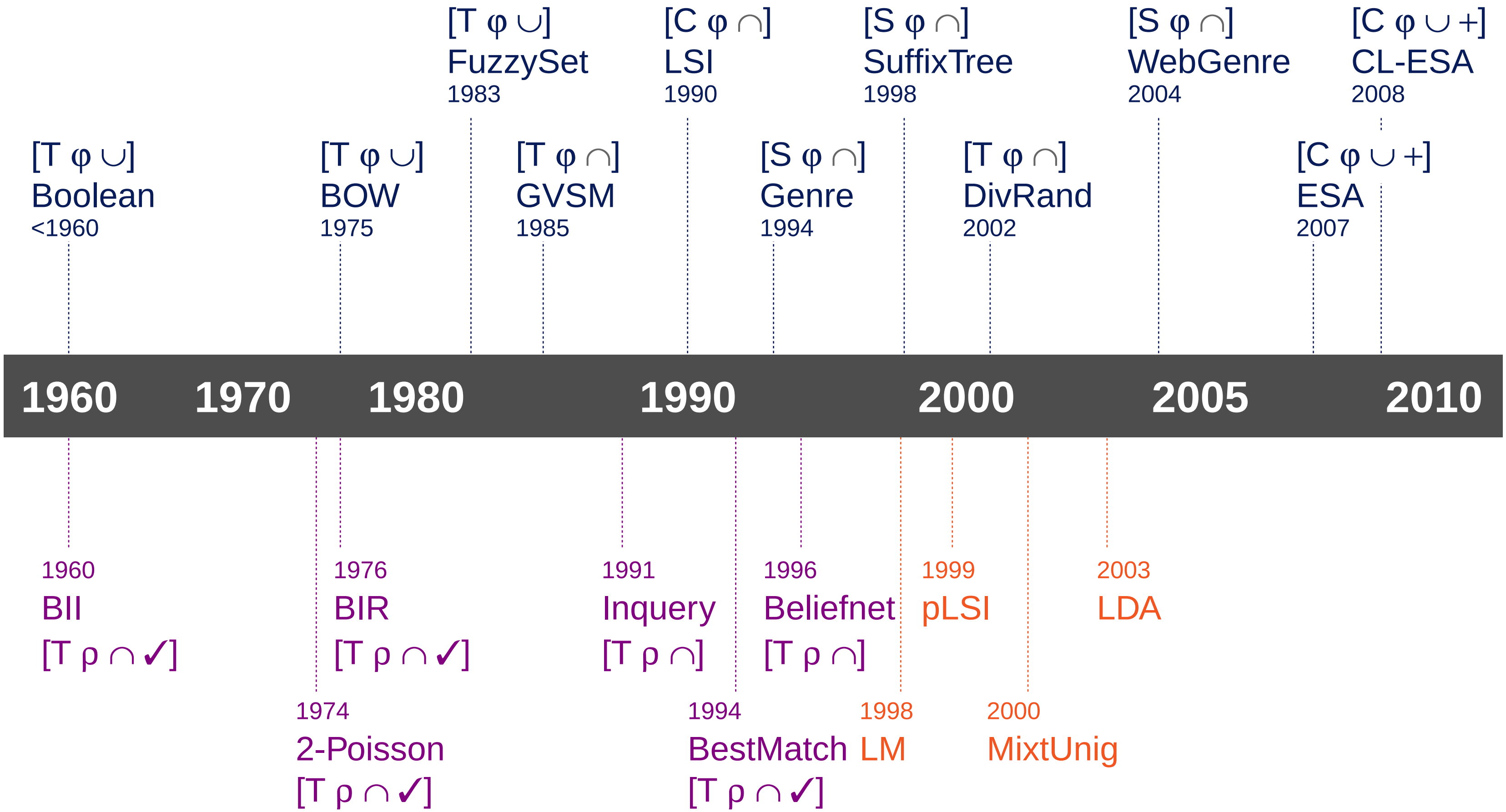
Retrieval Models



- Empirical models
- Probabilistic models
- Language models

Feature space	[T] terms [C] concepts [S] special
RSV foundation	[φ] sim. [ρ] relevance [γ] generation
Collection	[∪] open [∩] closed
Ext. knowledge	[✓] user feedback [+] collection

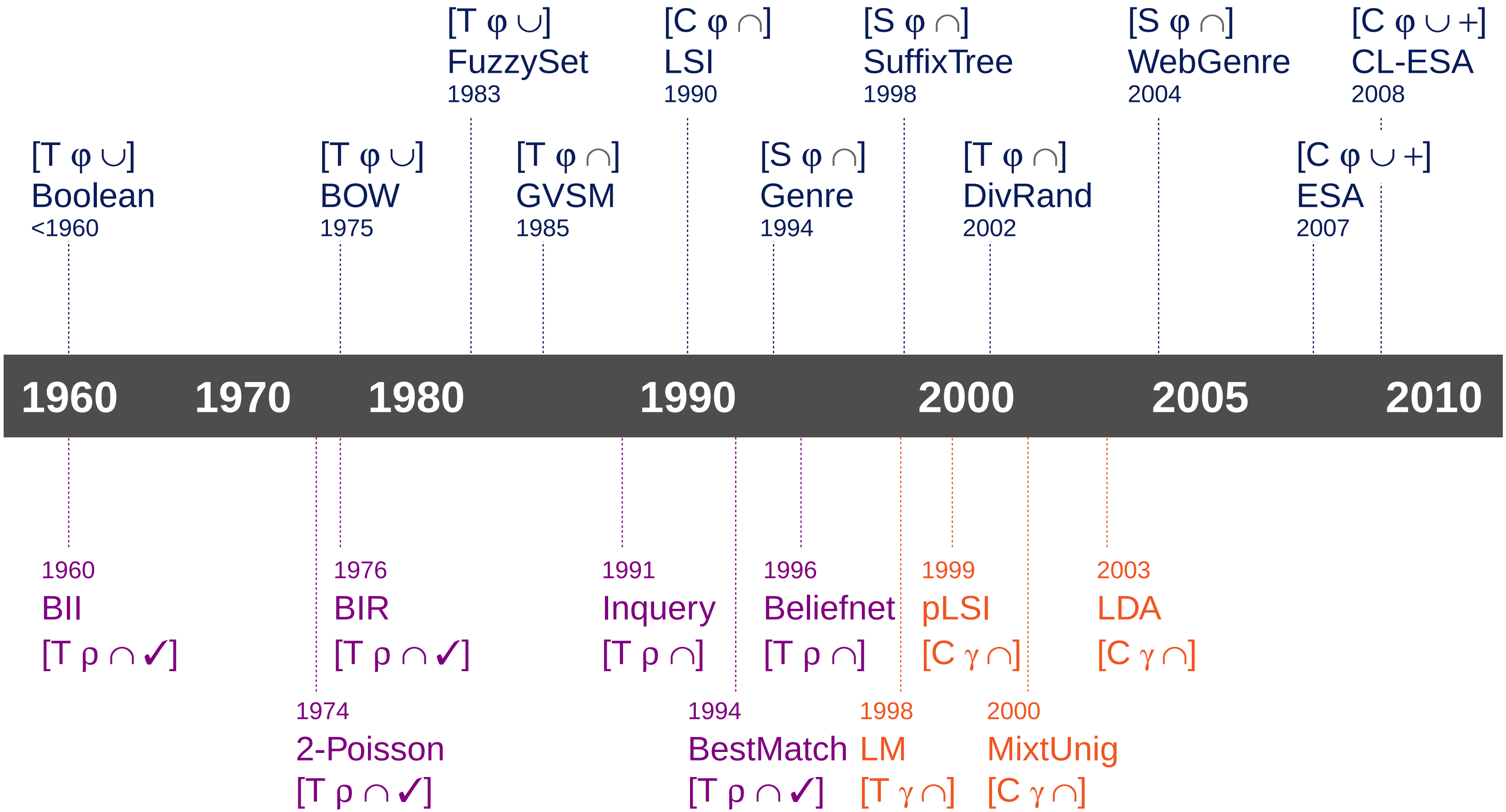
Retrieval Models



- Empirical models
- Probabilistic models
- Language models

Feature space	[T] terms [C] concepts [S] special
RSV foundation	[φ] sim. [ρ] relevance [γ] generation
Collection	[∪] open [∩] closed
Ext. knowledge	[✓] user feedback [+] collection

Retrieval Models



- Empirical models
- Probabilistic models
- Language models

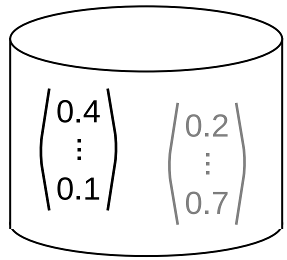
Feature space	[T] terms [C] concepts [S] special
RSV foundation	[φ] sim. [ρ] relevance [γ] generation
Collection	[∪] open [∩] closed
Ext. knowledge	[✓] user feedback [+] collection

The ESA Model Revisited

Explicit Semantic Analysis, ESA [Gabrilovich/Markovitch 2007]

The ESA Model Revisited

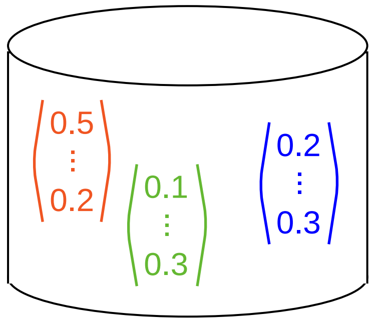
Explicit Semantic Analysis



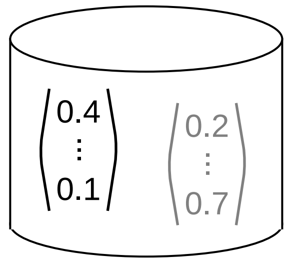
Document collection *D*

The ESA Model Revisited

Explicit Semantic Analysis



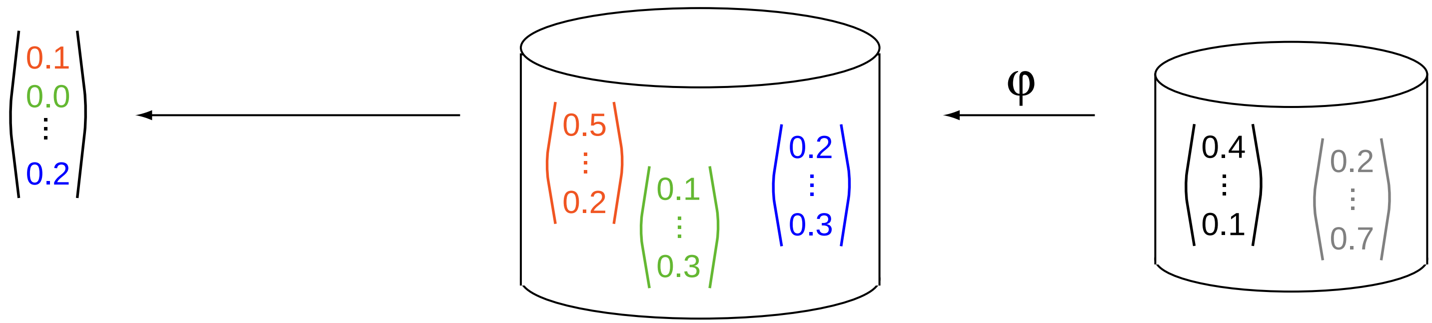
Index collection D_I
e.g. Wikipedia



Document
collection D

The ESA Model Revisited

Explicit Semantic Analysis



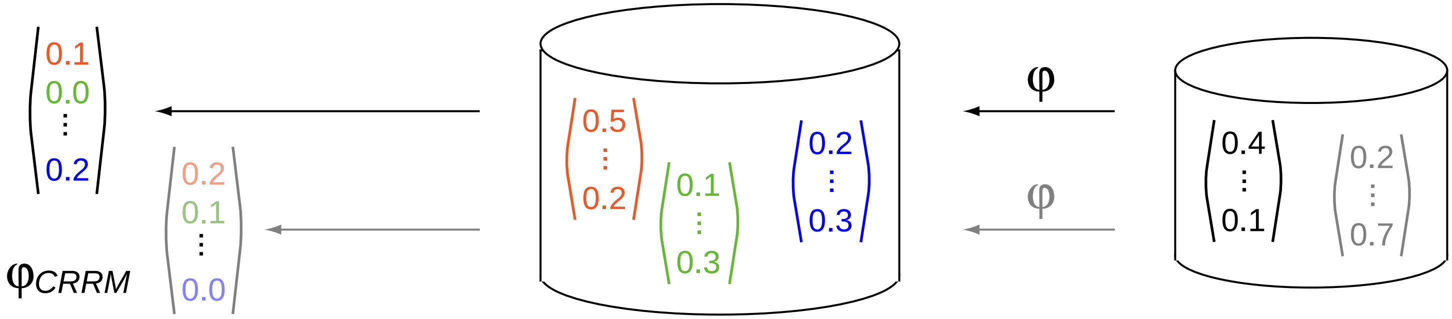
Concept space

Index collection D_I
e.g. Wikipedia

Document
collection D

The ESA Model Revisited

Explicit Semantic Analysis



Similarity analysis in a collection-relative Concept space

Index collection D_I e.g. Wikipedia

Document collection D

The ESA Model Revisited

ESA Characteristics [Gabrilovich/Markovitch, IJCAI 2007]

- representation is relative to an index collection, D_I, such as Wikipedia
- retrieval model to compute the semantic relatedness of text documents
- relies on robust IR technology: vector space model and cos-similarity
- document similarity under ESA entails substantial improvements

Retrieval model	**Correlation with human assessment**
Vector space model	0.50
Latent semantic indexing	0.60
ESA with Wikipedia	0.72
ESA with Open Directory Project	0.69

The ESA Model Revisited

ESA Characteristics [Gabrilovich/Markovitch, IJCAI 2007]

- representation is relative to an index collection, D_I, such as Wikipedia
- retrieval model to compute the semantic relatedness of text documents
- relies on robust IR technology: vector space model and cos-similarity
- document similarity under ESA entails substantial improvements

Retrieval model	**Correlation with human assessment**
Vector space model	0.50
Latent semantic indexing	0.60
ESA with Wikipedia	0.72
ESA with Open Directory Project	0.69

Experiment basis: 50 news documents from Australian Broadcasting Corporation.
1 225 human similarity assessments.

The ESA Model Revisited

Why or When does ESA Work?

The concept hypothesis:

- Each document in D_I describes exactly one concept.
- The concepts in D_I are “orthogonal”.

➜ D_I should provide some kind of “encyclopedic characteristic”

[Gabrilovich/Markovitch, IJCAI 2007]

The ESA Model Revisited

Why or When does ESA Work?

The concept hypothesis:

- Each document in D_I describes exactly one concept.
- The concepts in D_I are “orthogonal”.

➜ D_I should provide some kind of “encyclopedic characteristic”

[Gabrilovich/Markovitch, IJCAI 2007]

- The size (only) of D_I affects both accuracy and runtime of ESA.

➜ The concept hypothesis does not hold.

[Anderka/Stein, SIGIR 2009]

The ESA Model Revisited

ESA Evaluation [Anderka/Stein, SIGIR 2009]

Index collection	**Number of index documents**					
	1 000	10 000	50 000	100 000	150 000	200 000
VSM (baseline)	0.717	0.717	0.717	0.717	0.717	0.717
Wikipedia, *tf* · *idf*	0.742	**0.784**	0.782	0.782	0.781	0.781
Merged Topics, *tf* · *idf*	0.738	0.767	0.768	0.769	0.769	**0.777**
Reuters, *tf* · *idf*	0.767	0.795	**0.802**	0.800	0.800	0.800

Experiment basis: Retake of the Gabrilovich/Markovitch experiments.
Performance measured by Pearson's correlation coefficient.

The ESA Model Revisited

ESA Evaluation [Anderka/Stein, SIGIR 2009]

Index collection	**Number of index documents**					
	1 000	10 000	50 000	100 000	150 000	200 000
VSM (baseline)	0.717	0.717	0.717	0.717	0.717	0.717
Wikipedia, $tf \cdot idf$	0.742	**0.784**	0.782	0.782	0.781	0.781
Merged Topics, $tf \cdot idf$	0.738	0.767	0.768	0.769	0.769	**0.777**
Reuters, $tf \cdot idf$	0.767	0.795	**0.802**	0.800	0.800	0.800
Wikipedia, tf	0.704	0.724	0.732	0.732	**0.734**	0.732
Random Gaussian, tf	0.703	0.716	**0.717**	0.717	0.717	0.717

Experiment basis: Retake of the Gabrilovich/Markovitch experiments.
Performance measured by Pearson's correlation coefficient.

The ESA Model Revisited

ESA Evaluation [Anderka/Stein, SIGIR 2009]

Index collection	**Number of index documents**	
	1 000	10 000
VSM (baseline)	0.110	0.110
Wikipedia, $tf \cdot idf$	0.124	0.160
Merged Topics, $tf \cdot idf$	0.120	0.168
Reuters, $tf \cdot idf$	0.138	0.164

Experiment basis: 528 155 documents of the TREC-8 test collection.
Performance measured by Mean Average Precision, MAP.

The ESA Model Revisited

ESA Evaluation [Anderka/Stein, SIGIR 2009]

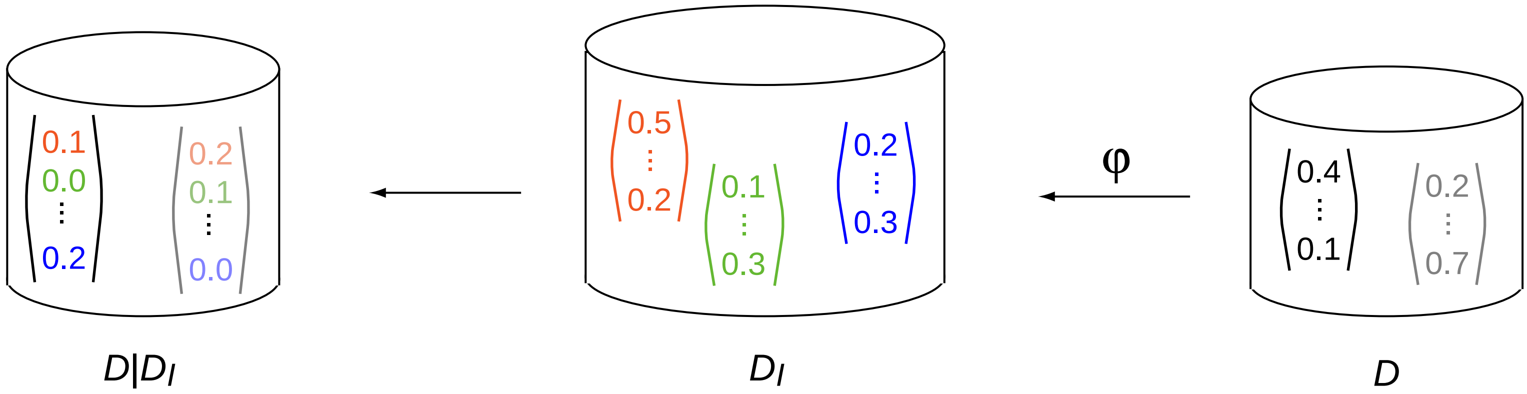
Index collection	**Number of index documents**	
	1 000	10 000
VSM (baseline)	0.110	0.110
Wikipedia, $tf \cdot idf$	0.124	0.160
Merged Topics, $tf \cdot idf$	0.120	0.168
Reuters, $tf \cdot idf$	0.138	0.164
Wikipedia, tf	0.111	0.141
Random Gaussian, tf	0.109	0.132

Experiment basis: 528 155 documents of the TREC-8 test collection.
Performance measured by Mean Average Precision, MAP.

Framework of Collection-Relative Retrieval Models

Framework of Collection-Relative Retrieval Models

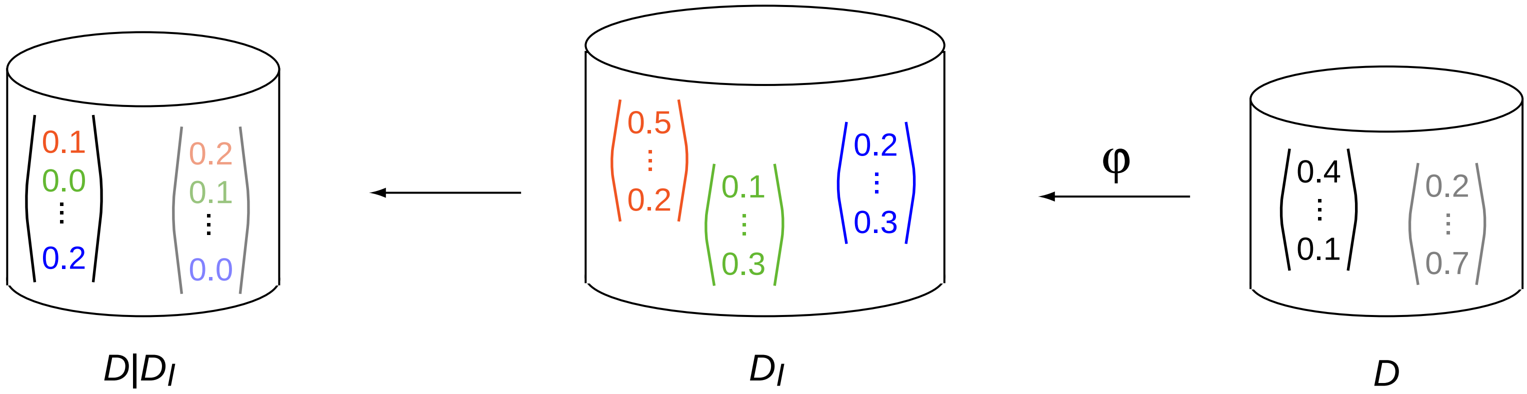
General



$$A_{D|D_I} = A_{D_I}^T \cdot A_D$$

Framework of Collection-Relative Retrieval Models

General



$$A_{D|D_I} = A_{D_I}^T \cdot A_D$$

Probability Ranking Principle:

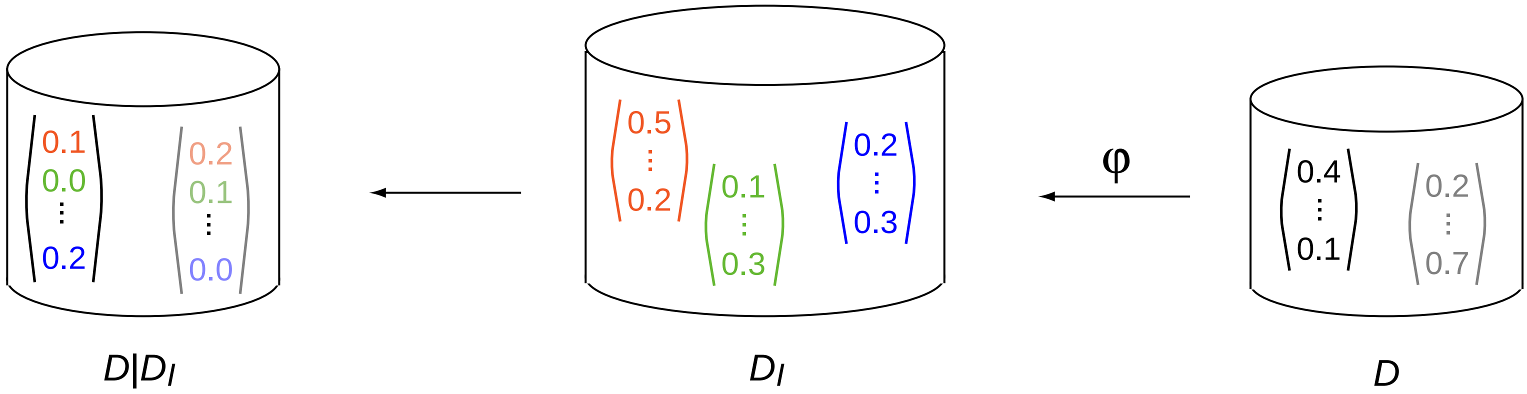
$$d^* = \text{argmax}_{d \in D} \ \varphi_{\mathit{CRRM}}(q, d),$$

where

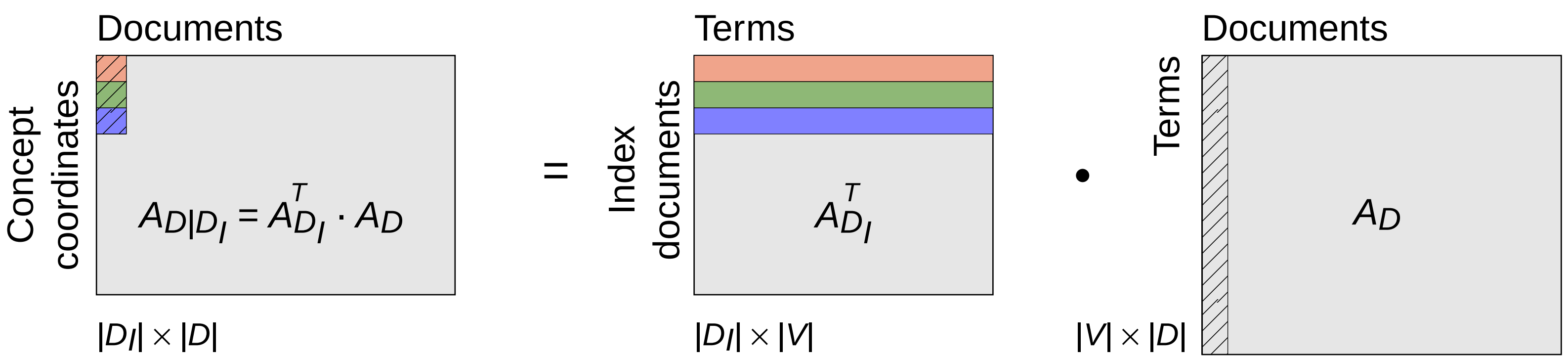
$$\varphi_{\mathit{CRRM}}(d_1, d_2) \ := \ \varphi(\mathbf{d}_{1|D_I}, \mathbf{d}_{2|D_I}) \ = \ \varphi(A_{D_I}^T \cdot \mathbf{d}_1, A_{D_I}^T \cdot \mathbf{d}_2)$$

Framework of Collection-Relative Retrieval Models

General

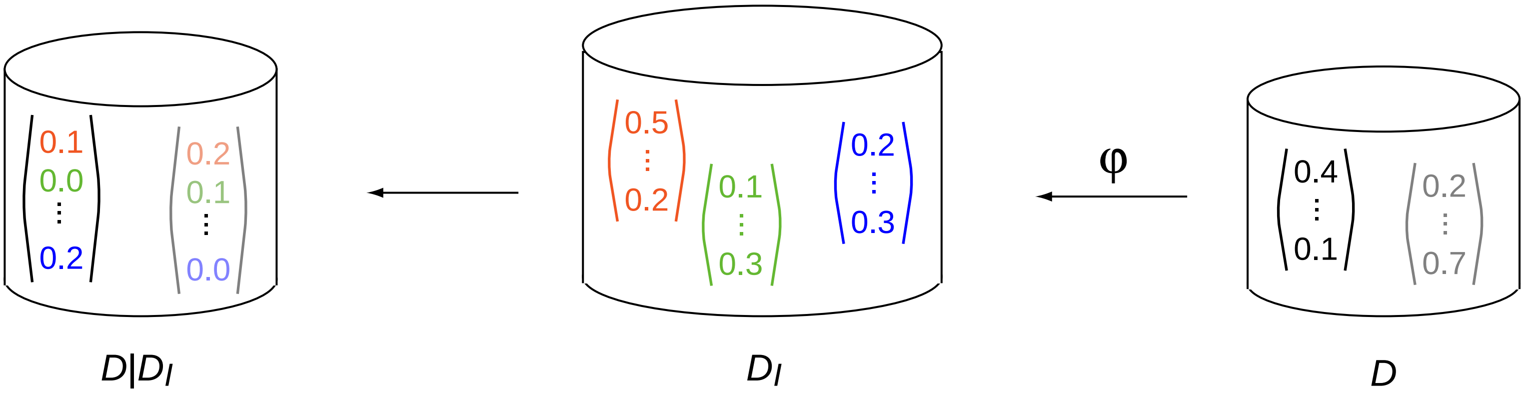


$$A_{D|D_I} = A_{D_I}^T \cdot A_D$$



Framework of Collection-Relative Retrieval Models

General



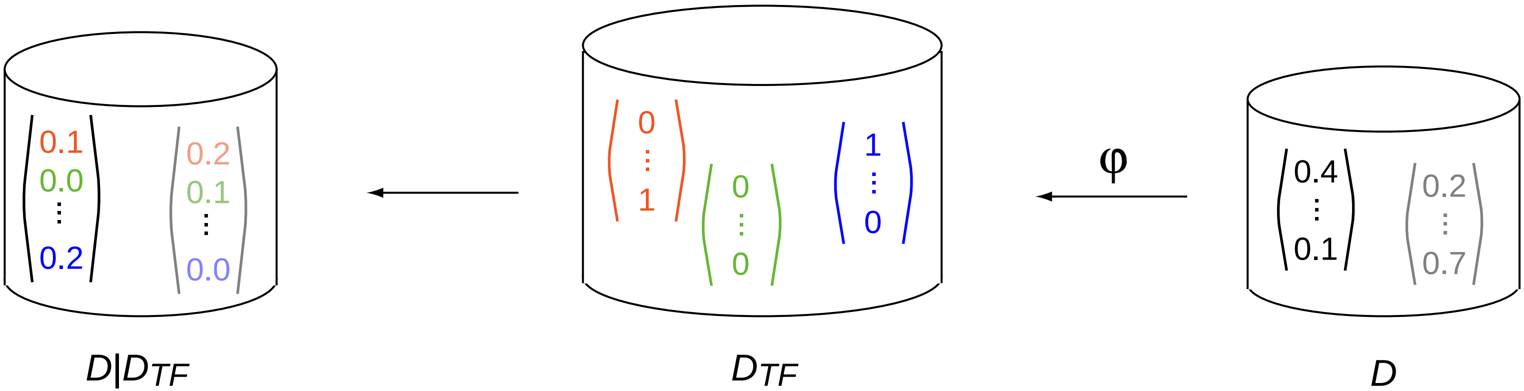
$$A_{D|D_I} = A_{D_I}^T \cdot A_D$$

CRRM taxonomy:

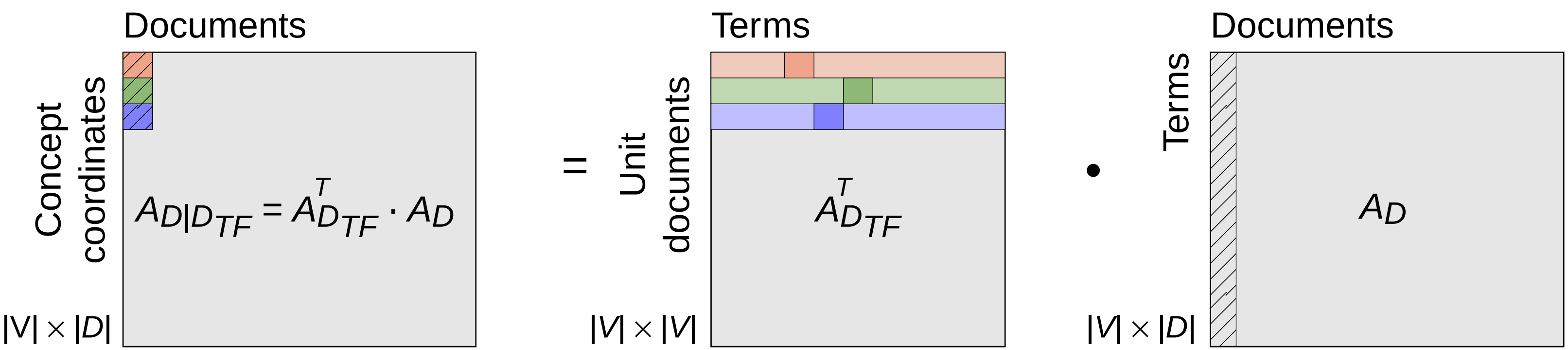
- Collection-relative retrieval model, CRRM
  - CRRM based on retrieval collection
    - VSM
    - GVSM
    - LSI
  - CRRM based on a foreign collection
    - ESA
    - CL-ESA
    - LSI (folding in)

Framework of Collection-Relative Retrieval Models

Vector Space Model

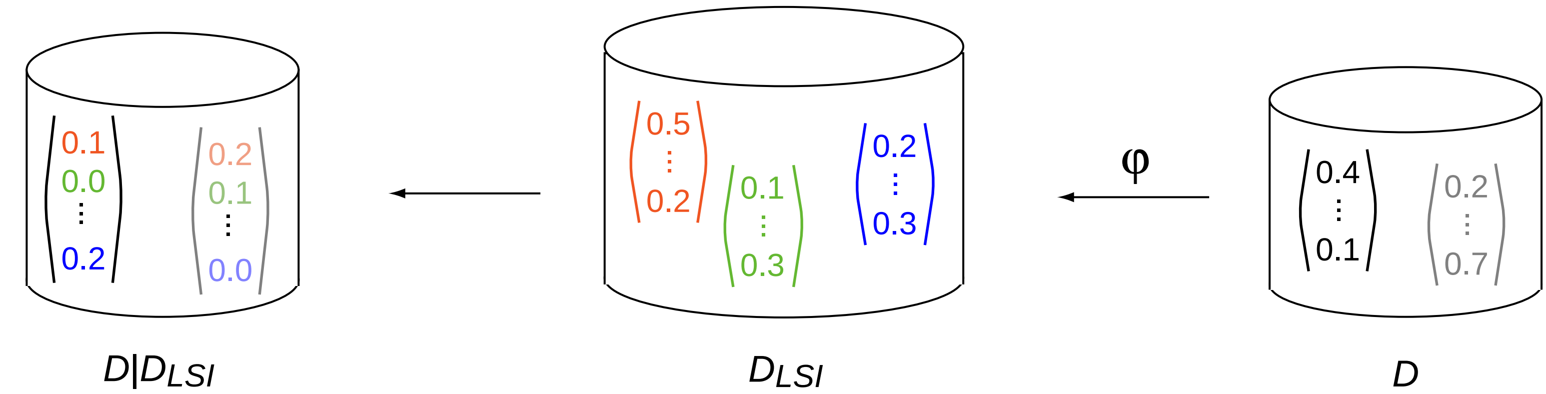


$$A_{D|D_{TF}} = A^T_{D_{TF}} \cdot A_D = A_D$$

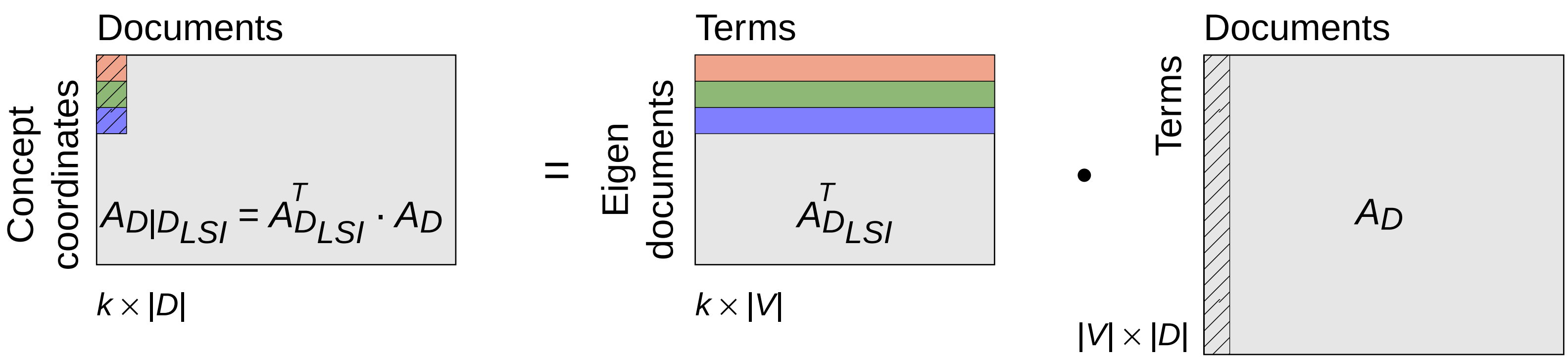


Framework of Collection-Relative Retrieval Models

Latent Semantic Indexing

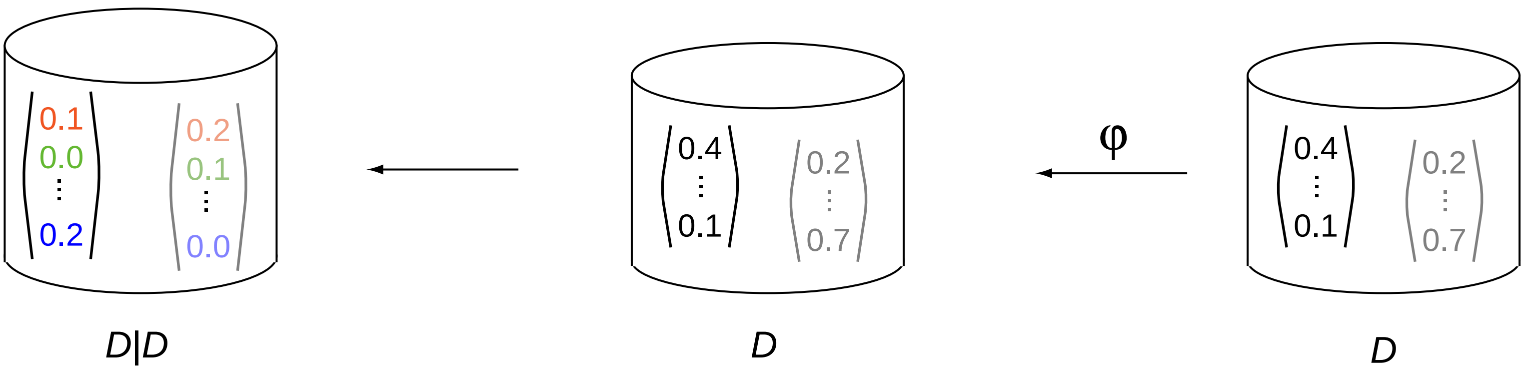


$$A_{D|D_{LSI}} = A^T_{D_{LSI}} \cdot A_D = \Sigma^{-1}_{D_k} \cdot U^T_{D_k} \cdot A_D$$

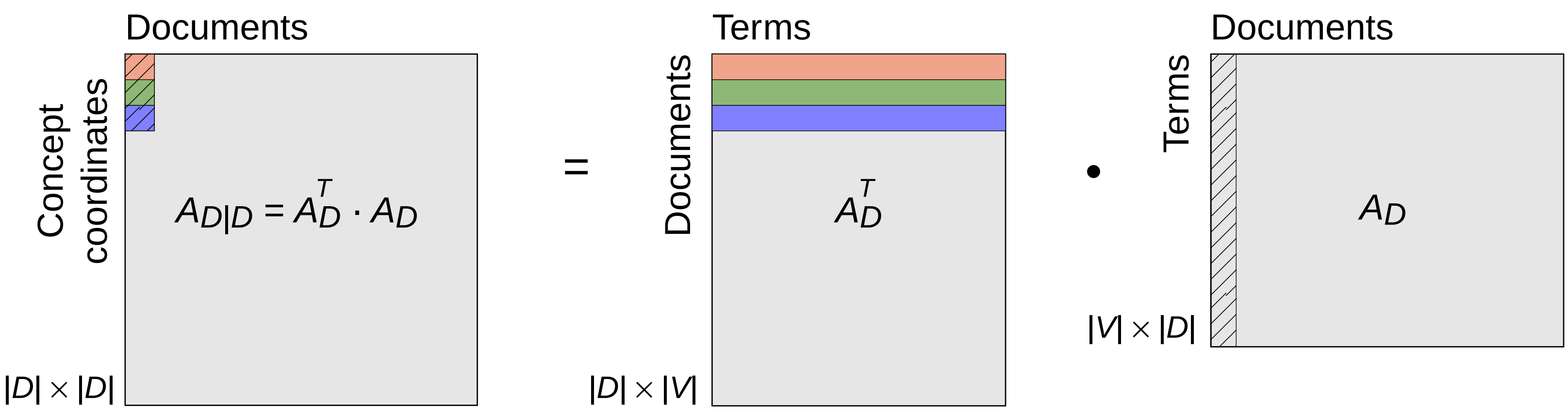


Framework of Collection-Relative Retrieval Models

Generalized Vector Space Model

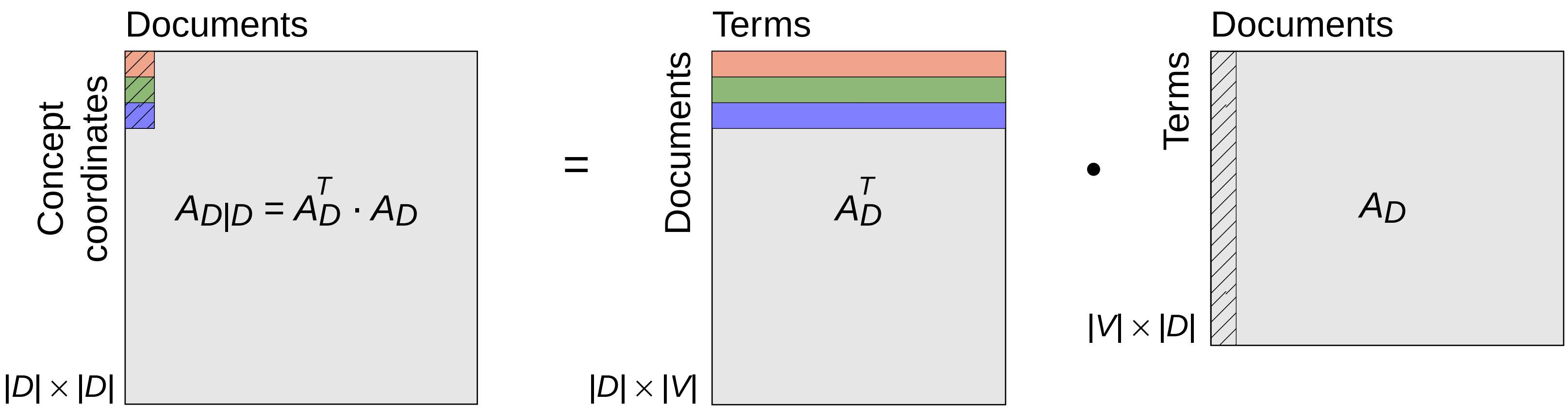


$$A_{D|D} = A_D^T \cdot A_D$$



Framework of Collection-Relative Retrieval Models

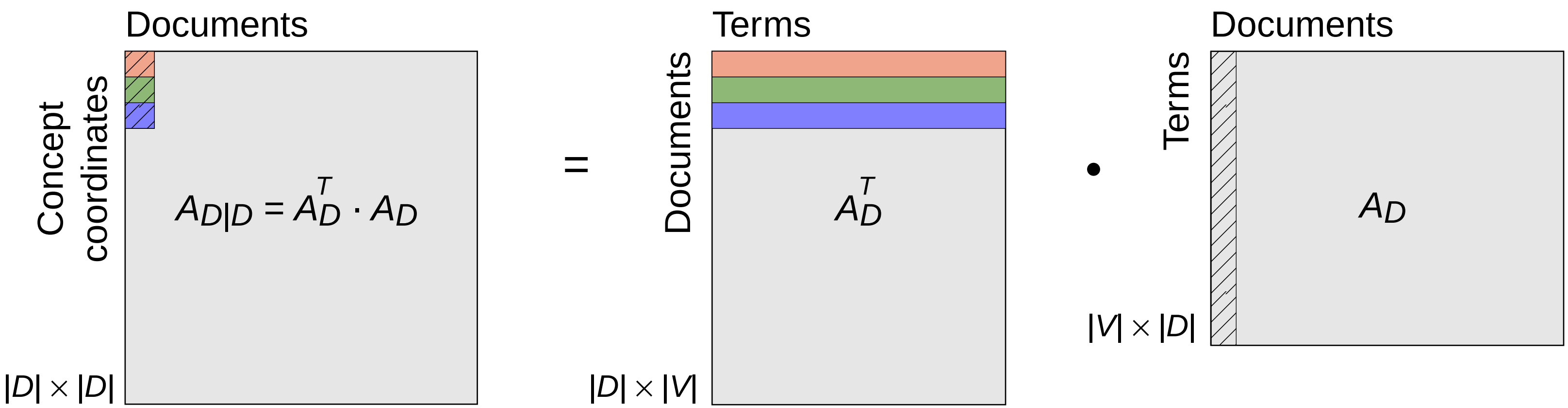
Generalized Vector Space Model (continued)



$$
\begin{aligned}
\varphi_{CRRM}(d_1, d_2) &= \varphi(\mathbf{d}_{1|D}, \mathbf{d}_{2|D}), \qquad \text{with } D_I := D \\
&= \varphi(A_D^T \cdot \mathbf{d}_1, A_D^T \cdot \mathbf{d}_2) \\
&= (A_D^T \cdot \mathbf{d}_1)^T \cdot A_D^T \cdot \mathbf{d}_2 \\
&= \mathbf{d}_1^T \cdot A_D \cdot A_D^T \cdot \mathbf{d}_2 \\
&= \mathbf{d}_1^T \cdot G \cdot \mathbf{d}_2 = \varphi_{GVSM}(d_1, d_2)
\end{aligned}
$$

Framework of Collection-Relative Retrieval Models

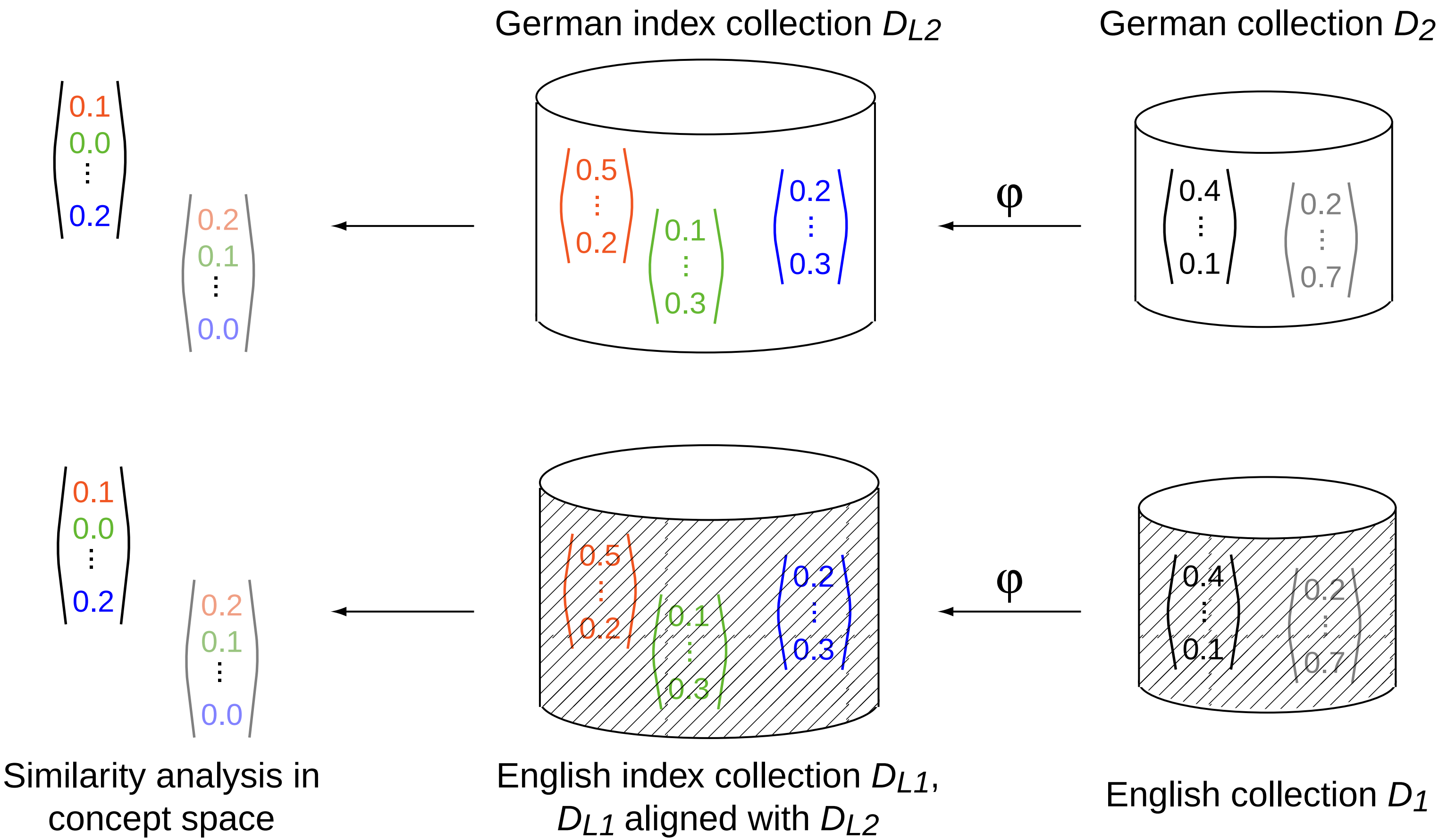
Generalized Vector Space Model (continued)



$$
\begin{aligned}
\varphi_{CRRM}(d_1, d_2) &= \varphi(\mathbf{d}_{1|D}, \mathbf{d}_{2|D}), \qquad \text{with } D_I := D \\
&= \varphi(A_D^T \cdot \mathbf{d}_1, A_D^T \cdot \mathbf{d}_2) \\
&= (A_D^T \cdot \mathbf{d}_1)^T \cdot A_D^T \cdot \mathbf{d}_2 \\
&= \mathbf{d}_1^T \cdot A_D \cdot A_D^T \cdot \mathbf{d}_2 \qquad \sim \text{Term co-occurrence} \\
&= \underbrace{\mathbf{d}_1^T \cdot G}_{\text{Query expansion}} \cdot \mathbf{d}_2 = \varphi_{GVSM}(d_1, d_2)
\end{aligned}
$$

Framework of Collection-Relative Retrieval Models

Cross Language ESA



$$d^* = \text{argmax}_{d \in D}\ \varphi(\mathbf{q}_{|D_{L_1}}, \mathbf{d}_{|D_{L_2}})$$

Framework of Collection-Relative Retrieval Models

Cross Language ESA (continued)

$$\begin{aligned}
\varphi_{\mathit{CRRM}}(d_1, d_2) &= \varphi(\mathbf{d}_{1|D_{L_1}}, \mathbf{d}_{2|D_{L_2}}), \qquad \text{with } D_{L_1}, D_{L_2} \text{ aligned} \\
&= \varphi(A^T_{D_{L_1}} \cdot \mathbf{d}_1, A^T_{D_{L_2}} \cdot \mathbf{d}_2) \\
&= (A^T_{D_{L_1}} \cdot \mathbf{d}_1)^T \cdot A^T_{D_{L_2}} \cdot \mathbf{d}_2 \\
&= \mathbf{d}_1^T \cdot A_{D_{L_1}} \cdot A^T_{D_{L_2}} \cdot \mathbf{d}_2 \qquad \sim \text{Cross language term co-occurrence} \\
&= \underbrace{\mathbf{d}_1^T \cdot G}_{\text{Query translation}} \cdot \mathbf{d}_2 = \varphi_{\mathit{GVSM}}(d_1, d_2)
\end{aligned}$$

Summary and Outlook

Summary and Outlook

Gabrilovich/Markovitch propose ESA in 2007.

They postulate the “concept hypothesis” to explain the success of ESA.

Summary and Outlook

Gabrilovich/Markovitch propose ESA in 2007.

They postulate the “concept hypothesis” to explain the success of ESA.

1. Concept hypothesis does not hold—size matters.
2. We abstract the ESA idea towards collection relativity.
   Well-known retrieval models can be understood as being collection-relative.
3. Query expansion under the GVSM $\equiv$ self collection relativity.
4. CL-ESA is a very powerful cross-language plagiarism detection technology.
   [Potthast/Stein/Anderka, ECIR 2008]
5. ESA may opens a new approach for specialized retrieval technology.
   ➜ Compilation of tailored index collections for special tasks.