TopicRank

Graph-Based Topic Ranking for Keyphrase Extraction

Adrien Bougouin Florian Boudin Béatrice Daille

Université de Nantes, LINA, France

16 October 2013

Introduction

Problem statement

Keyphrases

- Word or multi-word expressions
- **Overview** of a document's content

Applications

- Document indexing
- Document clustering
- Text summarization
- Query expansion
- Targeted advertising
- etc.

Lack of annotated documents

Many documents have **no associated keyphrases**.

Introduction

Problem statement

Keyphrases

- Word or multi-word expressions
- **Overview** of a document's content

Applications

- Document indexing
- Document clustering
- Text summarization
- Query expansion
- Targeted advertising
- etc.

Lack of annotated documents

Many documents have **no associated keyphrases**.

Introduction

Problem statement

Keyphrases

- Word or multi-word expressions
- **Overview** of a document's content

Applications

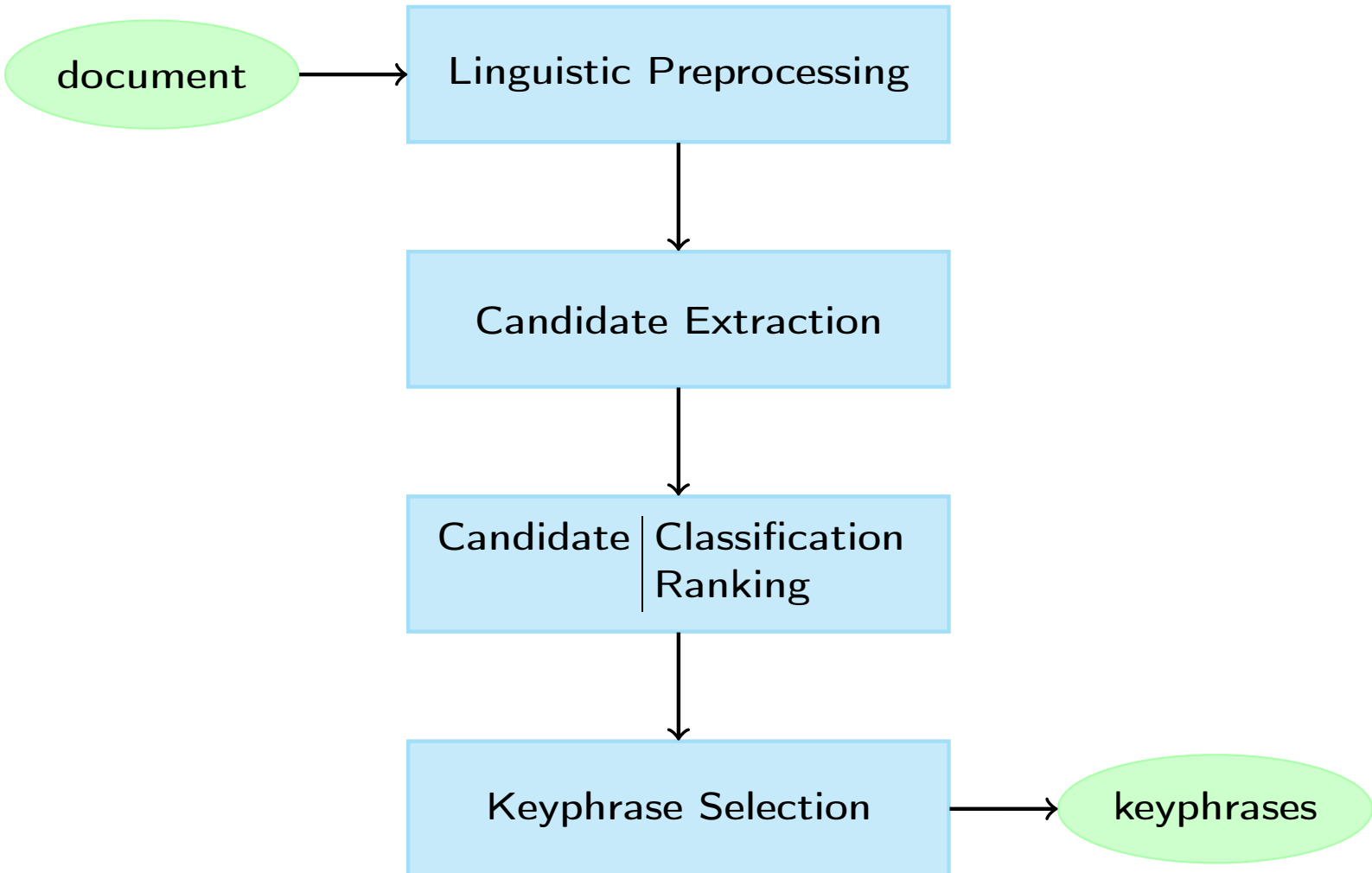
- Document indexing
- Document clustering
- Text summarization
- Query expansion
- Targeted advertising
- etc.

Lack of annotated documents

Many documents have **no associated keyphrases**.

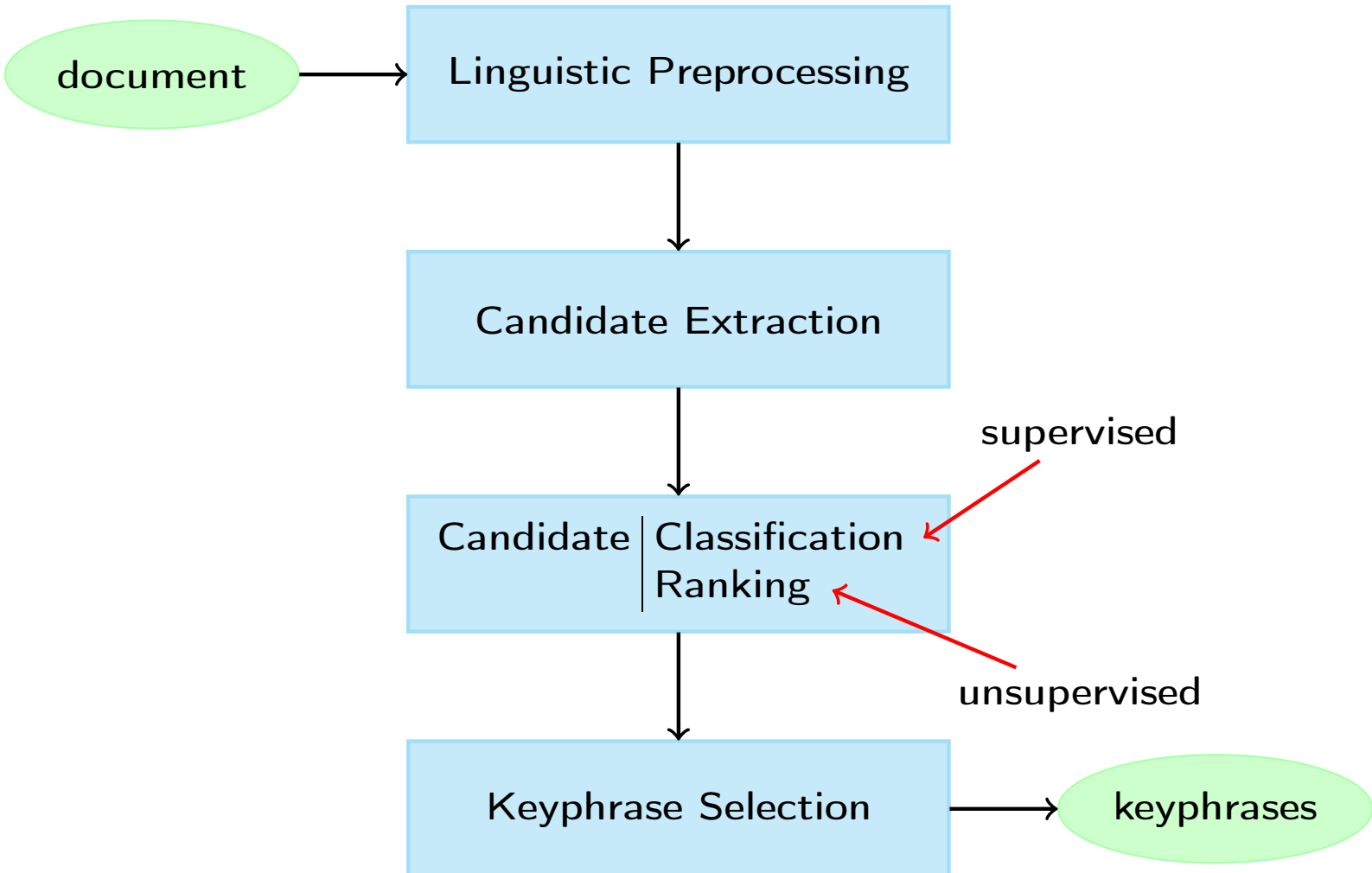
Introduction

Automatic keyphrase extraction



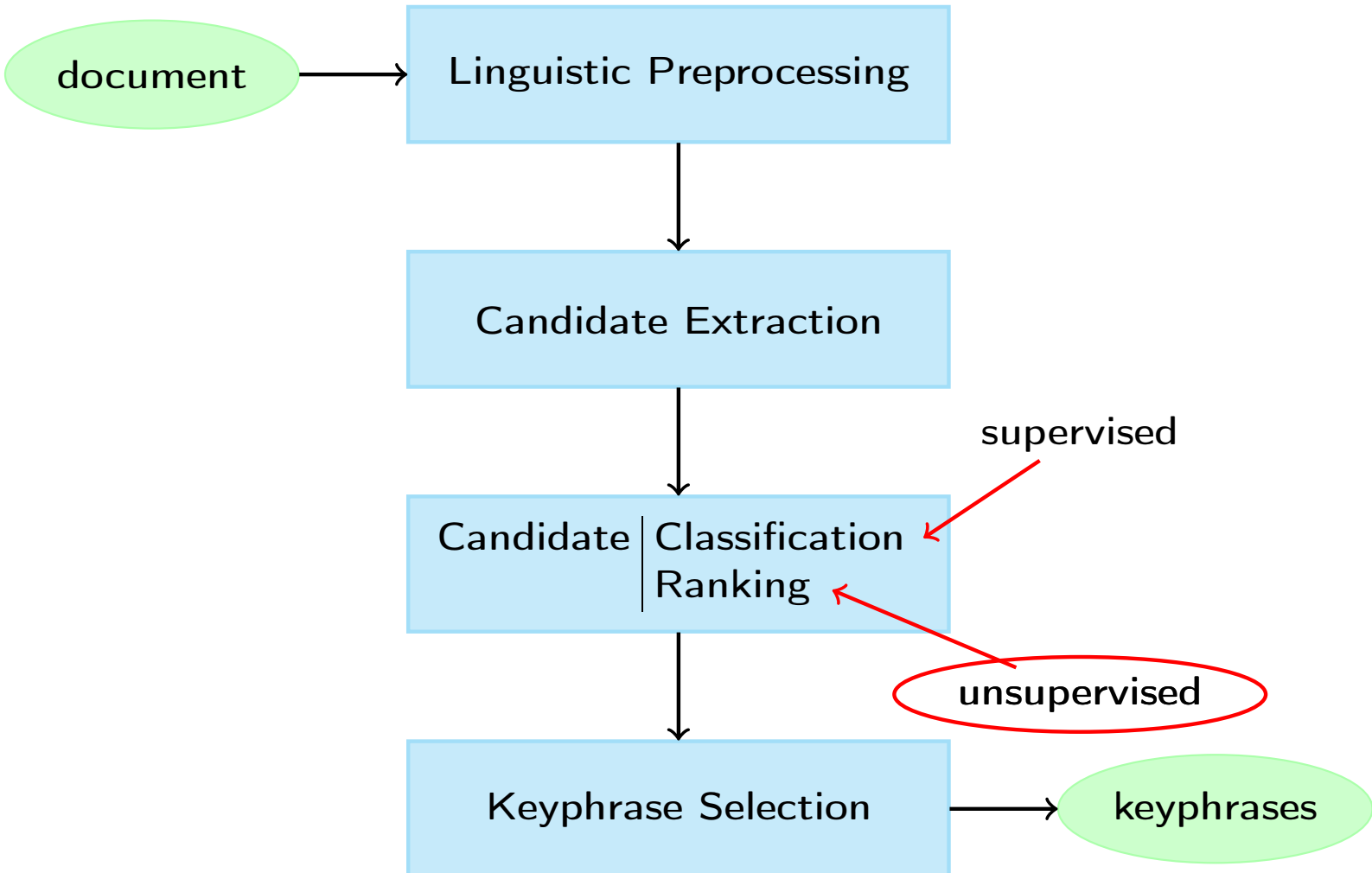
Introduction

Automatic keyphrase extraction



Introduction

Automatic keyphrase extraction



Introduction

Example

Project Euclid and the role of **research libraries** in **scholarly publishing**

Project Euclid, a **joint electronic journal publishing initiative** of **Cornell University Library** and **Duke University Press** is discussed in the broader contexts of the changing patterns of **scholarly communication** and the publishing scene of **mathematics**. Specific aspects of the project such as **partnerships** and the creation of an **economic model** are presented as well as what it takes to be a publisher. Libraries have gained important and relevant experience through the creation and management of digital libraries, but they need to develop further skills if they want to adopt a new role in the life cycle of **scholarly communication**.

Related Work

Unsupervised methods

Mostly ranking technics using:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measure
  - refined with similar documents
  - biased with topic probabilities

Related Work

Unsupervised methods

Mostly ranking technics using:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measure
  - refined with similar documents
  - biased with topic probabilities

(Tomokiyo and Hurst, 2003)

Related Work

Unsupervised methods

Mostly ranking technics using:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measure
  - refined with similar documents
  - biased with topic probabilities

(Liu et al., 2009)

Related Work

Unsupervised methods

Mostly ranking technics using:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measure
  - refined with similar documents
  - biased with topic probabilities

(Mihalcea and Tarau, 2004, TextRank)

Related Work

Unsupervised methods

Mostly ranking technics using:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measure
  - refined with similar documents
  - biased with topic probabilities

(Wan and Xiao, 2008; Tsatsaronis et al., 2010)

Related Work

Unsupervised methods

Mostly ranking technics using:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measure
  - refined with similar documents
  - biased with topic probabilities

(Wan and Xiao, 2008)

Related Work

Unsupervised methods

Mostly ranking technics using:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measure
  - refined with similar documents
  - biased with topic probabilities

(Liu et al., 2010)

Related Work

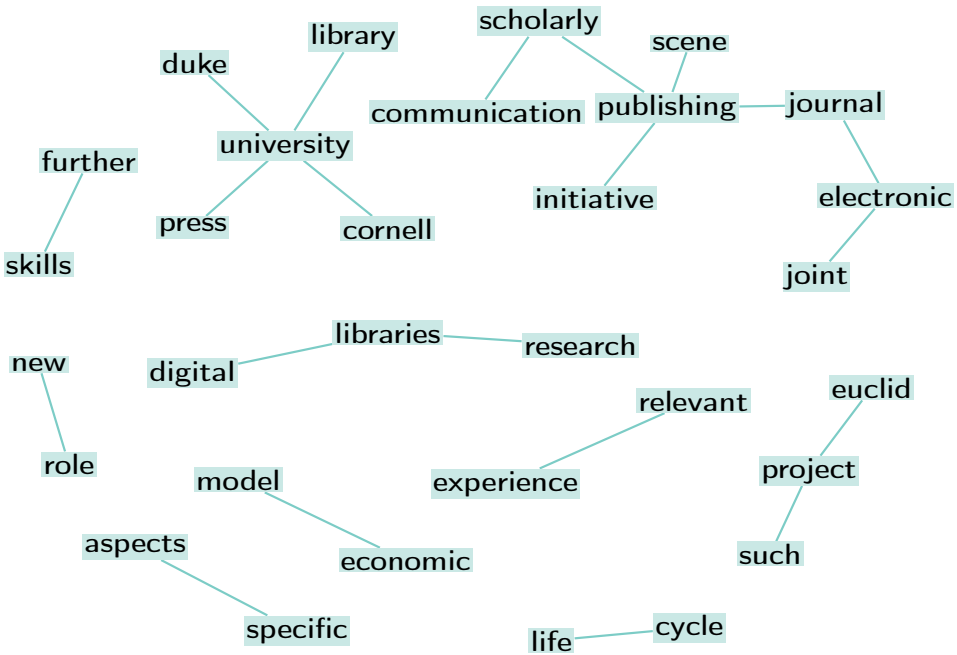
Graph-based approach: TextRank

Project Euclid and the role of research libraries in scholarly publishing

Project Euclid, a joint electronic journal publishing initiative of Cornell University Library and Duke University Press is discussed in the broader contexts of the changing patterns of scholarly communication and the publishing scene of mathematics. Specific aspects of the project such as partnerships and the creation of an economic model are presented as well as what it takes to be a publisher. Libraries have gained important and relevant experience through the creation and management of digital libraries, but they need to develop further skills if they want to adopt a new role in the life cycle of scholarly communication.

Related Work

Graph-based approach: TextRank



Generated Keyphrase

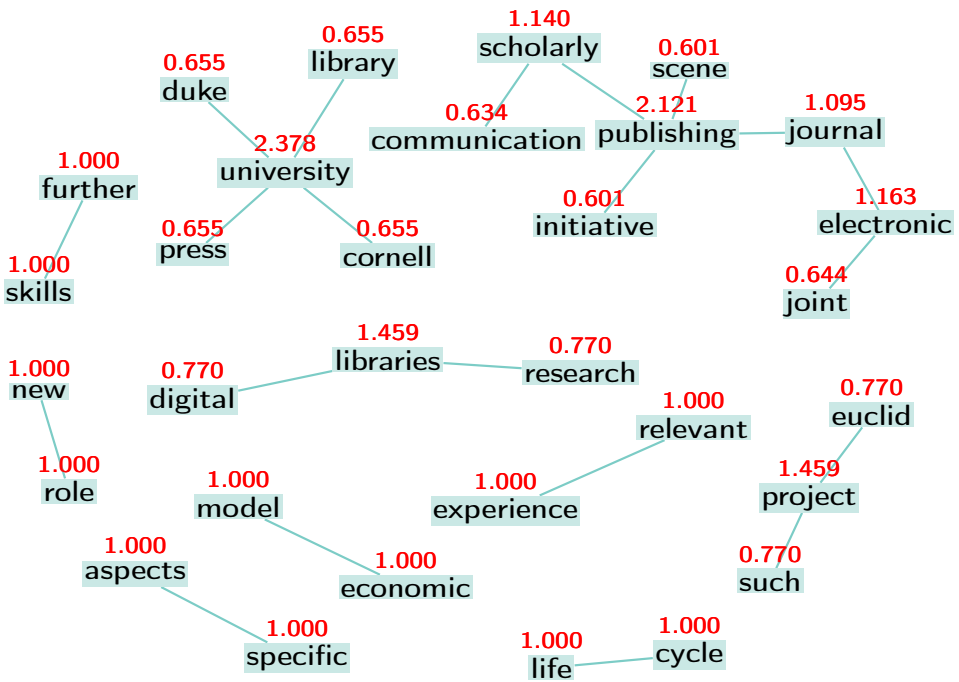
- electronic journal publishing
- scholarly publishing
- libraries
- university
- project
- economic
- relevant
- role

PageRank's "voting" concept

High-scoring words contribute more to the score of their connected words.

Related Work

Graph-based approach: TextRank



Generated Keyphrase

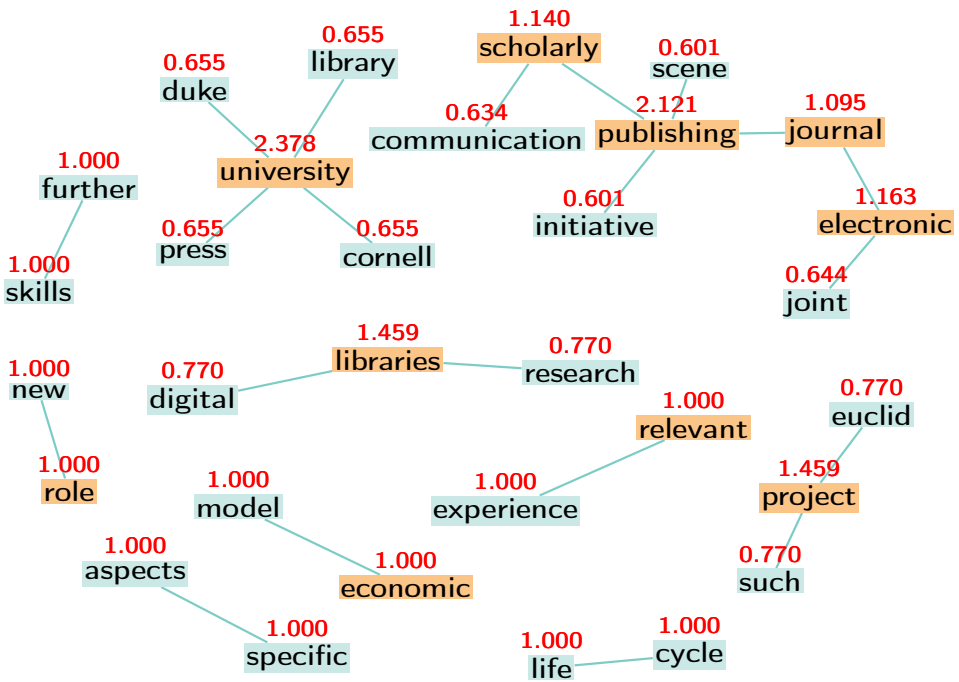
- electronic journal publishing
- scholarly publishing
- libraries
- university
- project
- economic
- relevant
- role

PageRank's "voting" concept

High-scoring words contribute more to the score of their connected words.

Related Work

Graph-based approach: TextRank



Generated Keyphrase

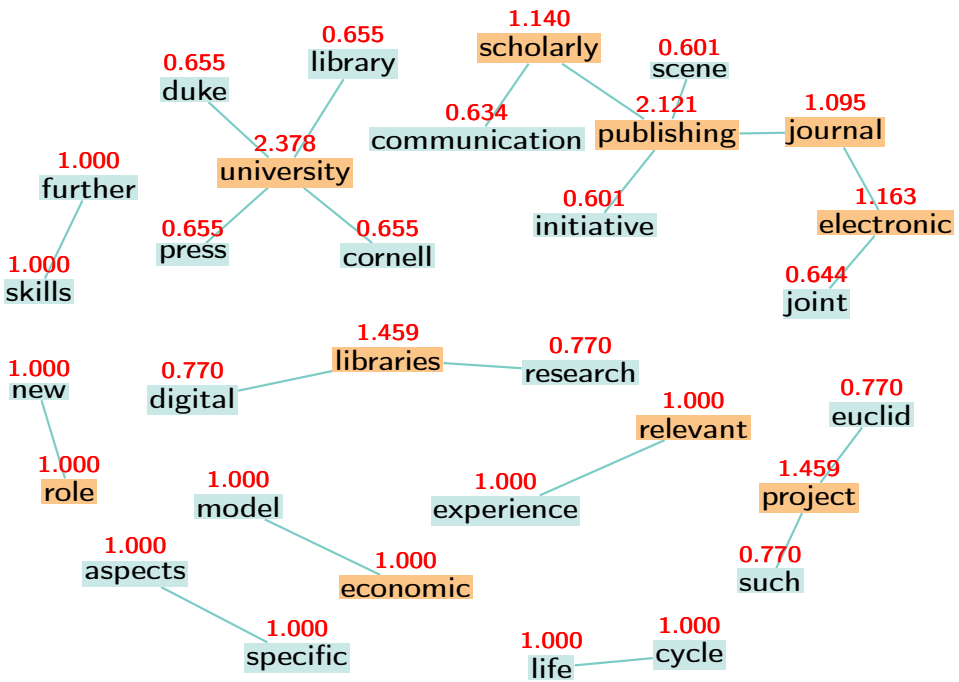
- electronic journal publishing
- scholarly publishing
- libraries
- university
- project
- economic
- relevant
- role

PageRank's "voting" concept

High-scoring words contribute more to the score of their connected words.

Related Work

Graph-based approach: TextRank



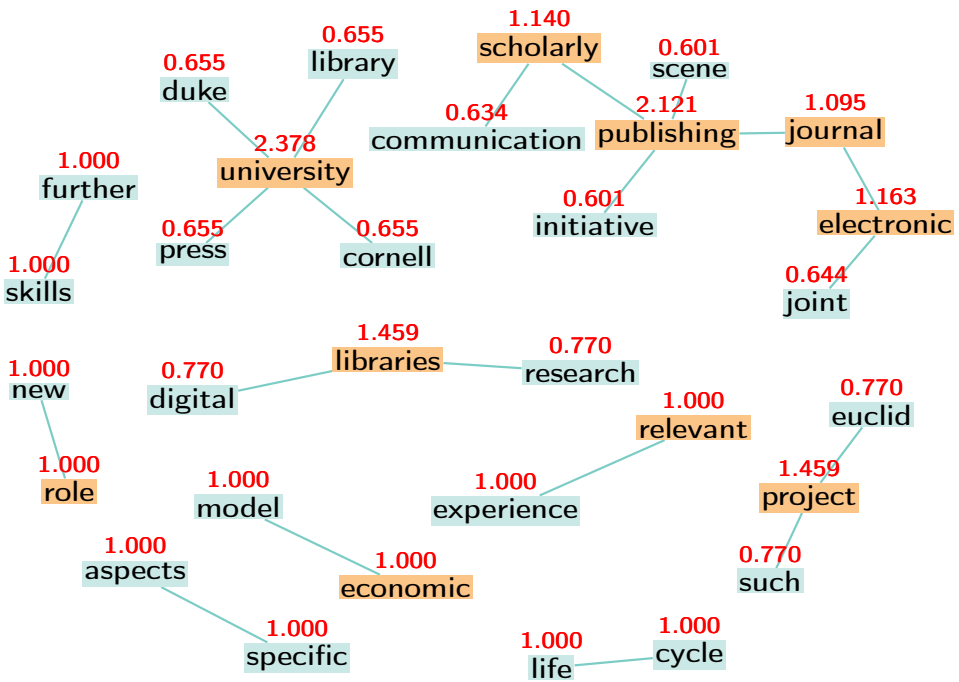
Generated Keyphrase
electronic journal publishing
scholarly publishing
libraries
university
project
economic
relevant
role

PageRank's "voting" concept

High-scoring words contribute more to the score of their connected words.

Related Work

Graph-based approach: TextRank



Generated Keyphrase
electronic journal publishing
scholarly publishing
libraries
university
project
economic
relevant
role

PageRank's "voting" concept

High-scoring words contribute more to the score of their connected words.

Related Work

Graph-based approach: TextRank

Limitations

- Word nodes
- Co-occurence window
- Several nodes for one topic

This Work

Limitations of previous work

- Word nodes
- Co-occurence window
- Several nodes for one topic

Proposal

1. Topic nodes
2. Complete graph construction

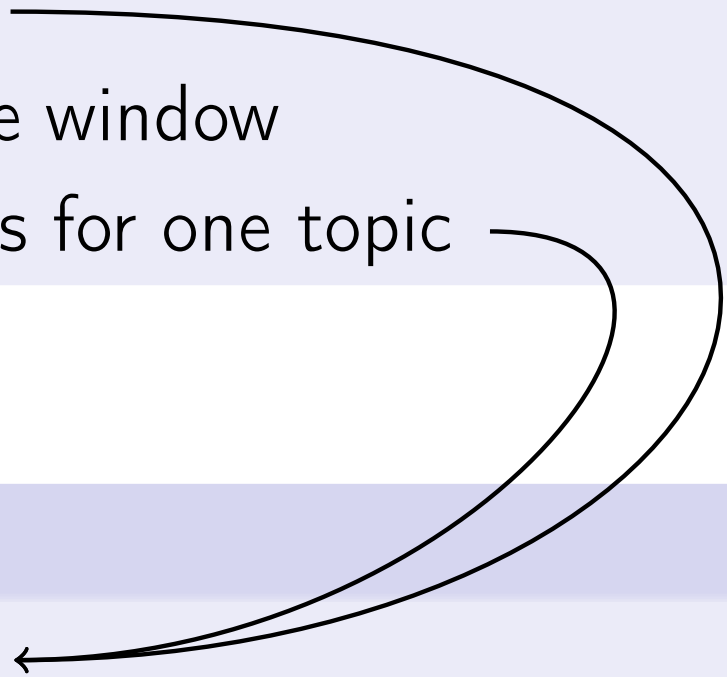
This Work

Limitations of previous work

- Word nodes
- Co-occurence window
- Several nodes for one topic

Proposal

1. Topic nodes
2. Complete graph construction



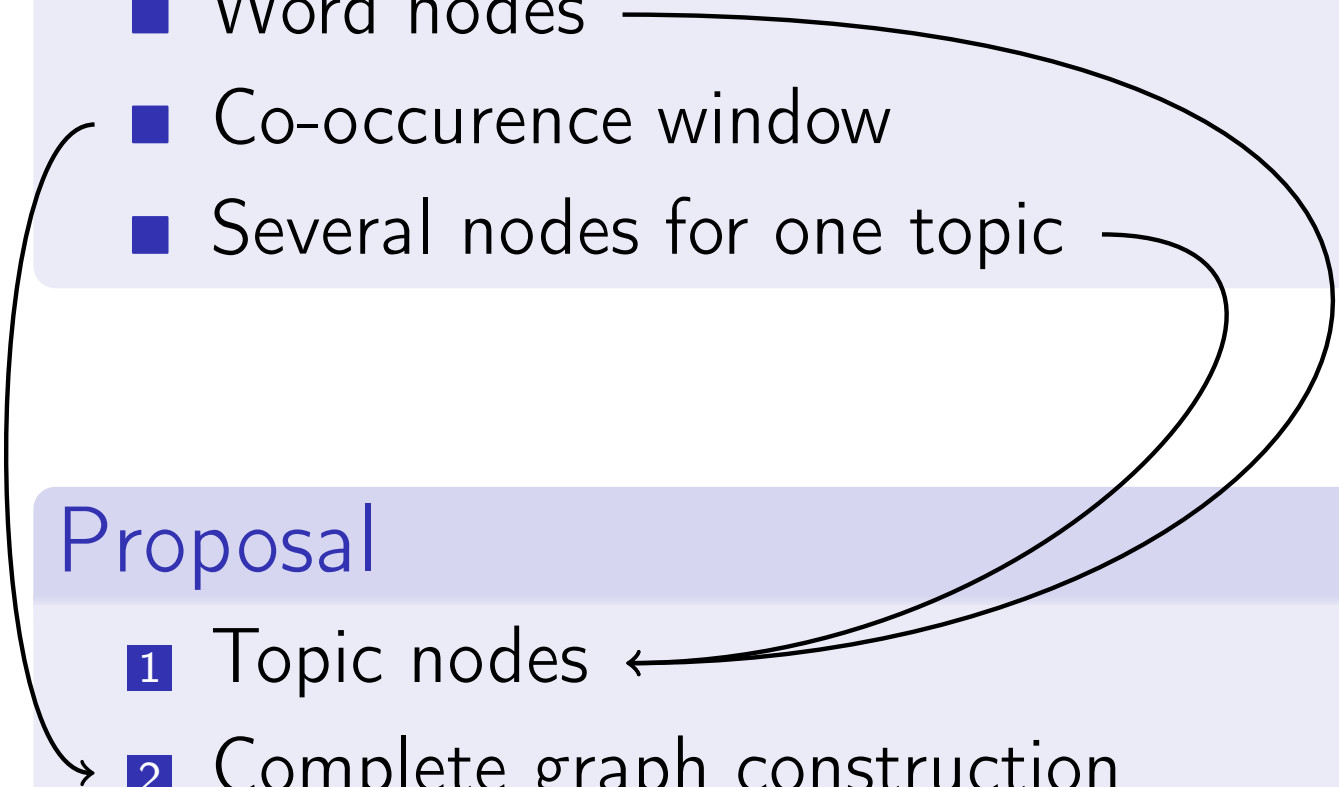
This Work

Limitations of previous work

- Word nodes
- Co-occurence window
- Several nodes for one topic

Proposal

1. Topic nodes
2. Complete graph construction



Plan

Plan

TopicRank

1. Candidate extraction
2. Candidate clustering
3. Graph construction
4. Topic ranking
5. Keyphrase selection

Project Euclid and the role of research libraries in scholarly publishing

Project Euclid, a joint electronic journal publishing initiative of Cornell University Library and Duke University Press is discussed in the broader contexts of the changing patterns of scholarly communication and the publishing scene of mathematics. Specific aspects of the project such as partnerships and the creation of an economic model are presented as well as what it takes to be a publisher. Libraries have gained important and relevant experience through the creation and management of digital libraries, but they need to develop further skills if they want to adopt a new role in the life cycle of scholarly communication.

TopicRank

1. Candidate extraction

⇒ (NOUN|ADJ)$^+$

2. Candidate clustering
3. Graph construction
4. Topic ranking
5. Keyphrase selection

Project Euclid and the role of research libraries in scholarly publishing

Project Euclid, a joint electronic journal publishing initiative of Cornell University Library and Duke University Press is discussed in the broader contexts of the changing patterns of scholarly communication and the publishing scene of mathematics. Specific aspects of the project such as partnerships and the creation of an economic model are presented as well as what it takes to be a publisher. Libraries have gained important and relevant experience through the creation and management of digital libraries, but they need to develop further skills if they want to adopt a new role in the life cycle of scholarly communication.

TopicRank

no linguistic knowledge

1. Candidate extraction

⇒ (NOUN|ADJ)$^+$

2. Candidate clustering
3. Graph construction
4. Topic ranking
5. Keyphrase selection

Project Euclid and the role of research libraries in scholarly publishing

Project Euclid, a joint electronic journal publishing initiative of Cornell University Library and Duke University Press is discussed in the broader contexts of the changing patterns of scholarly communication and the publishing scene of mathematics. Specific aspects of the project such as partnerships and the creation of an economic model are presented as well as what it takes to be a publisher. Libraries have gained important and relevant experience through the creation and management of digital libraries, but they need to develop further skills if they want to adopt a new role in the life cycle of scholarly communication.

TopicRank

1. Candidate extraction
2. Candidate clustering

⇒ Hierarchical clustering

3. Graph construction
4. Topic ranking
5. Keyphrase selection

ID	Topic
C01	cornell university library; digital libraries; research libraries; libraries
C02	project euclid; project such
C03	publishing scene; scholarly publishing; publisher
C04	role; new role $\longleftarrow$ stem overlap $\geq \frac{1}{4}$
C05	important
C06	scholarly communication
C07	further skills
C08	partnerships
C09	mathematics
C10	joint electronic journal publishing initiative
C11	contexts
C12	specific aspects
C13	economic model
C14	duke university press
C15	relevant experience
C16	creation
C17	life cycle
C18	patterns
C19	management

TopicRank

naive topic similarity

1. Candidate extraction
2. Candidate clustering

$\Rightarrow$ Hierarchical clustering

3. Graph construction
4. Topic ranking
5. Keyphrase selection

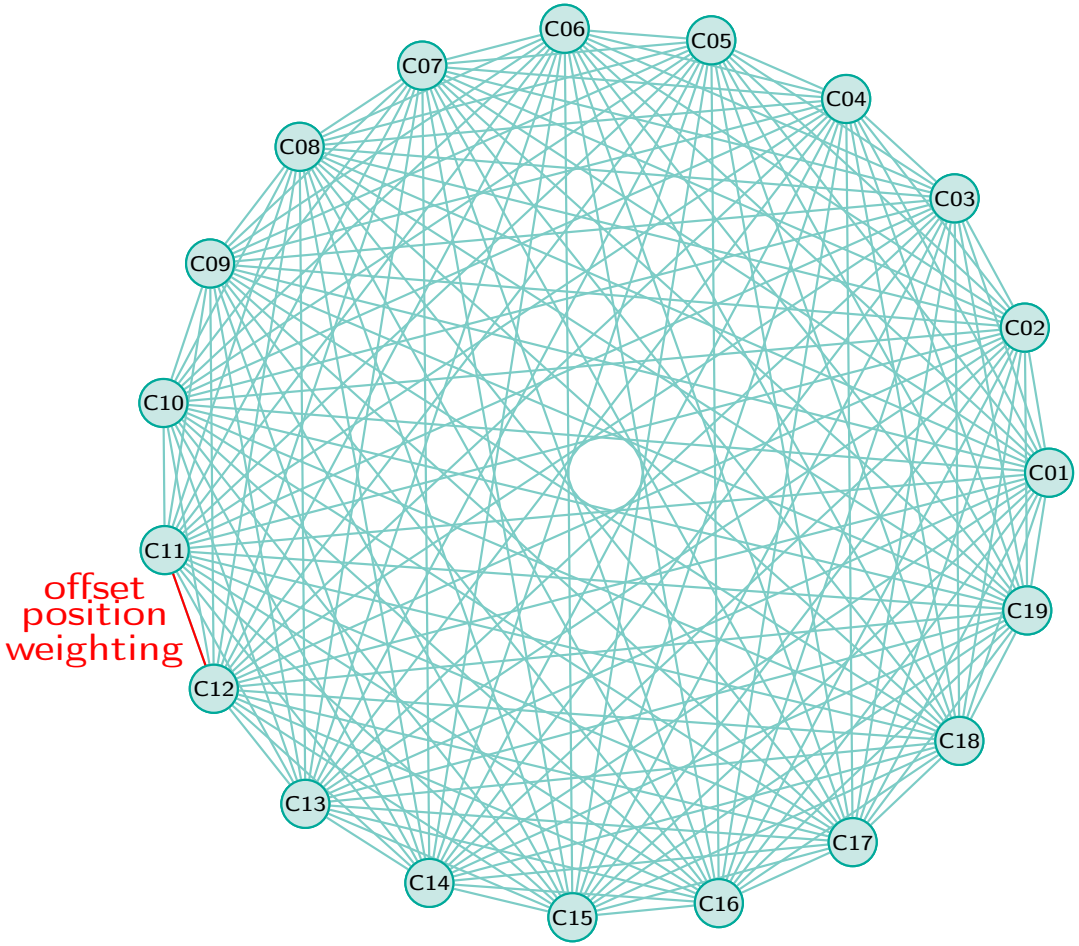
ID	Topic
C01	cornell university library; digital libraries; research libraries; libraries
C02	project euclid; project such
C03	publishing scene; scholarly publishing; publisher
C04	role; new role $\longleftarrow$ stem overlap $\geq \frac{1}{4}$
C05	important
C06	scholarly communication
C07	further skills
C08	partnerships
C09	mathematics
C10	joint electronic journal publishing initiative
C11	contexts
C12	specific aspects
C13	economic model
C14	duke university press
C15	relevant experience
C16	creation
C17	life cycle
C18	patterns
C19	management

TopicRank

1. Candidate extraction
2. Candidate clustering
3. Graph construction

$\Rightarrow$ Complete graph

4. Topic ranking
5. Keyphrase selection

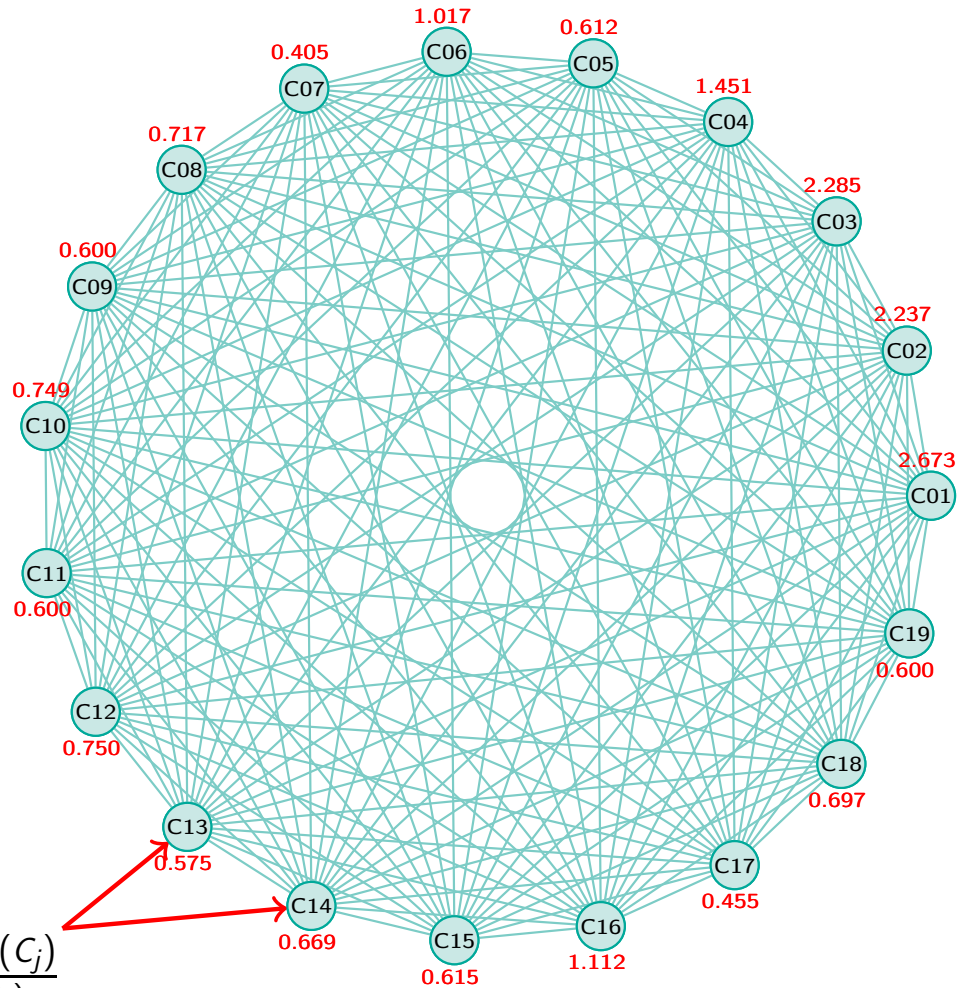


TopicRank

1. Candidate extraction
2. Candidate clustering
3. Graph construction
4. Topic ranking

$\Rightarrow$ PageRank's scoring

5. Keyphrase selection



$$\text{score}(C_i) = (1-\lambda) + \lambda \times \sum_{C_j \neq C_i} \frac{\text{weight}(C_j, C_i) \times \text{score}(C_j)}{\sum_{C_k \neq C_j} \text{weight}(C_j, C_k)}$$

TopicRank

1. Candidate extraction
2. Candidate clustering
3. Graph construction
4. Topic ranking
5. Keyphrase selection

$\Rightarrow$ First appearing one

Rank	ID	Topic
01	C01	cornell university library; digital libraries; research libraries; libraries
02	C03	publishing scene; scholarly publishing; publisher
03	C02	project euclid; project such
04	C04	role; new role
05	C16	creation
06	C06	scholarly communication
07	C09	mathematics
08	C12	specific aspects
09	C10	joint electronic journal publishing initiative
10	C08	partnerships
...	...	

TopicRank

Project Euclid and the role of research libraries in scholarly publishing

Project Euclid, a joint electronic journal publishing initiative of Cornell University Library and Duke University Press is discussed in the broader contexts of the changing patterns of scholarly communication and the publishing scene of mathematics. [...] Libraries have gained important and relevant experience through the creation and management of digital libraries, but they need to develop further skills if they want to adopt a new role in the life cycle of scholarly communication.

Rank	ID	Topic
01	C01	cornell university library; digital libraries; research libraries; libraries
02	C03	publishing scene; scholarly publishing; publisher
03	C02	project euclid; project such
04	C04	role; new role
05	C16	creation
06	C06	scholarly communication
07	C09	mathematics
08	C12	specific aspects
09	C10	joint electronic journal publishing initiative
10	C08	partnerships
...	...	

TopicRank

1. Candidate extraction
2. Candidate clustering
3. Graph construction
4. Topic ranking
5. Keyphrase selection

$\Rightarrow$ First appearing one

Rank	ID	Topic
01	C01	cornell university library; digital libraries; research libraries; libraries
02	C03	publishing scene; scholarly publishing; publisher
03	C02	project euclid; project such
04	C04	role; new role
05	C16	creation
06	C06	scholarly communication
07	C09	mathematics
08	C12	specific aspects
09	C10	joint electronic journal publishing initiative
10	C08	partnerships
...	...	

TopicRank

Project Euclid and the role of research libraries in scholarly publishing

Project Euclid, a joint electronic journal publishing initiative of Cornell University Library and Duke University Press is discussed in the broader contexts of the changing patterns of scholarly communication and the publishing scene of mathematics. Specific aspects of the project such as partnerships and the creation of an economic model are presented as well as what it takes to be a publisher. [...]

Rank	ID	Topic
01	C01	cornell university library; digital libraries; research libraries; libraries
02	C03	publishing scene; scholarly publishing; publisher
03	C02	project euclid; project such
04	C04	role; new role
05	C16	creation
06	C06	scholarly communication
07	C09	mathematics
08	C12	specific aspects
09	C10	joint electronic journal publishing initiative
10	C08	partnerships
...	...	

TopicRank

1. Candidate extraction
2. Candidate clustering
3. Graph construction
4. Topic ranking
5. Keyphrase selection

$\Rightarrow$ First appearing one

Rank	ID	Topic
01	C01	cornell university library; digital libraries; research libraries; libraries
02	C03	publishing scene; scholarly publishing; publisher
03	C02	project euclid; project such
04	C04	role; new role
05	C16	creation
06	C06	scholarly communication
07	C09	mathematics
08	C12	specific aspects
09	C10	joint electronic journal publishing initiative
10	C08	partnerships
...	...	

TopicRank

Project Euclid and the role of research libraries in scholarly publishing

[...] Specific aspects of the project such as partnerships and the creation of an economic model are presented as well as what it takes to be a publisher. [...]

Rank	ID	Topic
01	C01	cornell university library; digital libraries; research libraries; libraries
02	C03	publishing scene; scholarly publishing; publisher
03	C02	project euclid; project such
04	C04	role; new role
05	C16	creation
06	C06	scholarly communication
07	C09	mathematics
08	C12	specific aspects
09	C10	joint electronic journal publishing initiative
10	C08	partnerships
...	...	

TopicRank

1. Candidate extraction
2. Candidate clustering
3. Graph construction
4. Topic ranking
5. Keyphrase selection

$\Rightarrow$ First appearing one

Rank	ID	Topic
01	C01	cornell university library; digital libraries; research libraries; libraries
02	C03	publishing scene; scholarly publishing; publisher
03	C02	project euclid; project such
04	C04	role; new role
05	C16	creation
06	C06	scholarly communication
07	C09	mathematics
08	C12	specific aspects
09	C10	joint electronic journal publishing initiative
10	C08	partnerships
...	...	

TopicRank

Project Euclid and the role of research libraries in scholarly publishing

[...] Libraries have gained important and relevant experience through the creation and management of digital libraries, but they need to develop further skills if they want to adopt a new role in the life cycle of scholarly communication.

Rank	ID	Topic
01	C01	cornell university library; digital libraries; research libraries; libraries
02	C03	publishing scene; scholarly publishing; publisher
03	C02	project euclid; project such
04	C04	role; new role
05	C16	creation
06	C06	scholarly communication
07	C09	mathematics
08	C12	specific aspects
09	C10	joint electronic journal publishing initiative
10	C08	partnerships
...	...	

TopicRank

1. Candidate extraction
2. Candidate clustering
3. Graph construction
4. Topic ranking
5. Keyphrase selection

$\Rightarrow$ First appearing one

Rank	ID	Topic
01	C01	cornell university library; digital libraries; research libraries; libraries
02	C03	publishing scene; scholarly publishing; publisher
03	C02	project euclid; project such
04	C04	role; new role
05	C16	creation
06	C06	scholarly communication
07	C09	mathematics
08	C12	specific aspects
09	C10	joint electronic journal publishing initiative
10	C08	partnerships
...	...	

TopicRank

1. Candidate extraction
2. Candidate clustering
3. Graph construction
4. Topic ranking
5. Keyphrase selection

$\Rightarrow$ First appearing one

Rank	ID	Topic
01	C01	cornell university library; digital libraries; research libraries; libraries
02	C03	publishing scene; scholarly publishing; publisher
03	C02	project euclid; project such
04	C04	role; new role
05	C16	creation
06	C06	scholarly communication
07	C09	mathematics
08	C12	specific aspects
09	C10	joint electronic journal publishing initiative
10	C08	partnerships
...	...	

TopicRank

1. Candidate extraction
2. Candidate clustering
3. Graph construction
4. Topic ranking
5. Keyphrase selection

$\Rightarrow$ First appearing one

Keyphrase
research libraries
scholarly publishing
project euclid
role
creation
scholarly communication
mathematics
specific aspects
joint electronic journal publishing initiative
partnerships
...

TopicRank

1. Candidate extraction
2. Candidate clustering
3. Graph construction
4. Topic ranking
5. Keyphrase selection

$\Rightarrow$ First appearing one

Keyphrase
research libraries
scholarly publishing
project euclid
role
creation
scholarly communication
mathematics
specific aspects
joint electronic journal publishing initiative
partnerships
...

Plan

Evaluation

Datasets

Two English datasets:

- Inspec contains 500 abstracts of journal papers
  - 136.3 tokens/document
- SemEval (2010) contains 100 scientific papers
  - 5179.6 tokens/document

Two French datasets:

- WikiNews contains 100 news articles
  - 309.6 tokens/document
- DEFT (2012) contains 93 scientific papers
  - 6844.0 tokens/document

Evaluation

Datasets

Two English datasets:

- Inspec contains 500 abstracts of journal papers
  - 136.3 tokens/document
- SemEval (2010) contains 100 scientific papers
  - 5179.6 tokens/document

Two French datasets:

- WikiNews contains 100 news articles
  - 309.6 tokens/document
- DEFT (2012) contains 93 scientific papers
  - 6844.0 tokens/document

Evaluation

Datasets

Two English datasets:

- Inspec contains 500 abstracts of journal papers
  - 136.3 tokens/document
- SemEval (2010) contains 100 scientific papers
  - 5179.6 tokens/document

Two French datasets:

- WikiNews contains 100 news articles
  - 309.6 tokens/document
- DEFT (2012) contains 93 scientific papers
  - 6844.0 tokens/document

Evaluation

Datasets

Two English datasets:

- Inspec contains 500 abstracts of journal papers
  - 136.3 tokens/document
- SemEval (2010) contains 100 scientific papers
  - 5179.6 tokens/document

Two French datasets:

- WikiNews contains 100 news articles
  - 309.6 tokens/document
- DEFT (2012) contains 93 scientific papers
  - 6844.0 tokens/document

Evaluation

Datasets

Two English datasets:

- Inspec contains 500 abstracts of journal papers
  - 136.3 tokens/document
- SemEval (2010) contains 100 scientific papers
  - 5179.6 tokens/document

Two French datasets:

- WikiNews contains 100 news articles
  - 309.6 tokens/document
- DEFT (2012) contains 93 scientific papers
  - 6844.0 tokens/document

Evaluation

Baselines

- TF-IDF weighting
- TextRank
  - Word co-occurrence graph with a window of 2
  - Keyphrase generation based on keywords (10-bests)
- SingleRank
  - Word co-occurrence graph with a window of 10
  - Candidate keyphrases scored by their words' score (sum)

Evaluation

Baselines

- TF-IDF weighting
- TextRank
  - Word co-occurrence graph with a window of 2
  - Keyphrase generation based on keywords (10-bests)
- SingleRank
  - Word co-occurrence graph with a window of 10
  - Candidate keyphrases scored by their words' score (sum)

Evaluation

Baselines

- TF-IDF weighting
- TextRank
  - Word co-occurrence graph with a window of 2
  - Keyphrase generation based on keywords (10-bests)
- SingleRank
  - Word co-occurrence graph with a window of 10
  - Candidate keyphrases scored by their words' score (sum)

Evaluation

Measures

- Cut-off at 10 keyphrases
- F-score $\Rightarrow$ compromise between precision and recall

$$\text{f-score} = (1+\beta^2) \times \frac{\text{precision} \times \text{recall}}{(\beta^2 \times \textit{precision}) + \textit{recall}}$$

$$\beta = 1$$

- Problem of dealing with gold standard

$\Rightarrow$ Stemmed form comparisons

Evaluation

Main results

Method	Inspec	SemEval	WikiNews	DEFT
TF-IDF	33.4	10.5	34.3	13.2
TextRank	12.7	5.6	8.6	5.7
SingleRank	35.2	3.7	19.7	5.9
TopicRank	27.9	12.1	35.6	15.1

- Improvement over TF-IDF
- Significant improvement over graph-based methods
- Performance loss on Inspec

Evaluation

Individual contributions

Method	Inspec	SemEval	WikiNews	DEFT
SingleRank	35.2	3.7	19.7	5.9
+phrases	22.1	8.0	28.9	13.5
+topics	26.8	11.9	31.4	14.8
+complete	35.5	4.4	20.3	5.8
TopicRank	27.9	12.1	35.6	15.1

- Nodes: Topics $>$ candidates $>$ words
- Complete graph $\geq$ co-occurrence graph
- Contribution improve performances
- The above statements are false on Inspec

Evaluation

Keyphrase selection

Keyphrase selection	Inspec	SemEval	WikiNews	DEFT
First position	27.9	12.1	35.6	15.1
Frequency	26.8	1.4	26.2	2.5
Centroid	24.7	1.5	28.5	3.4
Upper bound	35.6	30.3	42.9	19.3

- Still room for improvement

Plan

Conclusion and Future Work

What we have done:

- Proposed TopicRank
- Topic ranking instead of word ranking
- Complete graph
- Experiments conducted of four standard datasets
- Good results
- Promising upper bound results

Still to do:

- Experiment various topic identifications
- Provide a keyphrase selection strategy getting closer to the upper bound

Conclusion and Future Work

What we have done:

- Proposed TopicRank
- Topic ranking instead of word ranking
- Complete graph
- Experiments conducted of four standard datasets
- Good results
- Promising upper bound results

Still to do:

- Experiment various topic identifications
- Provide a keyphrase selection strategy getting closer to the upper bound

Thank you

Backups

Candidate Extraction

- Focusing on nouns and adjectives is "enough" for English
- Prepositions and determiners should also be considered for French

Statistic	Corpus	
	SemEval	DEFT
Containing nouns	95.9%	79.3%
Containing proper nouns	5.8%	16.8%
Containing adjectives	40.5%	28.8%
Containing verbs	3.4%	0.5%
Containing adverbs	0.6%	0.5%
Containing prepositions	1.2%	12.7%
Containing determiners	0.0%	8.1%
Containing others	2.1%	5.8%

Backups

Candidate Clustering

The hierarchical clustering is an iterative algorithm:

- Initial state: candidates keyphrases are clusters
- Clusters with the highest similarity are merged together
- Clusters similarity is the average similarity between their candidates c_i:

$$\text{similarity}(c_1, c_2) = \frac{||\text{stem}(c_1) \cap \text{stem}(c_2)||}{||\text{stem}(c_1) \cup \text{stem}(c_2)||}$$

- A similarity threshold is set to 0.25

Backups

Graph Construction

- Nodes are topics
- Every nodes are connected to each other
- Connections between topics are weighted by the semantic strength between them
- Topics appearing close to each other have a high semantic strength:

$$\text{weight}(t_i, t_j) = \sum_{c_i \in t_i} \sum_{c_j \in t_j} \text{dist}(c_i, c_j)$$

$$\text{dist}(c_i, c_j) = \sum_{p_i \in \text{pos}(c_i)} \sum_{p_j \in \text{pos}(c_j)} \frac{1}{|p_i - p_j|}$$

Backups

Graph Construction

	Inspec	**SemEval**	**WikiNews**	**DEFT**
clusters/documents	20.9	272.4	52.4	546.5

Backups

Topic Ranking

PageRank's "voting" concept

High-scoring topics contribute more to the score of their connected topics.

$$\text{score}(C_i) = (1-\lambda) + \lambda \times \sum_{C_j \neq C_i} \frac{\text{weight}(C_i, C_j) \times \text{score}(C_j)}{\sum_{C_k \neq C_j} \text{weight}(C_j, C_k)}$$

$$\lambda = 0.85$$

Backups

Main Results

Method	Inspec			SemEval			WikiNews			DEFT		
	P	R	F	P	R	F	P	R	F	P	R	F
TF-IDF	32.7	38.6	33.4	13.2	8.9	10.5	33.9	35.9	34.3	10.3	19.1	13.2
TextRank	14.2	12.5	12.7	7.9	4.5	5.6	9.3	8.3	8.6	4.9	7.1	5.7
SingleRank	34.8	40.4	35.2	4.6	3.2	3.7	19.4	20.7	19.7	4.5	9.0	5.9
TopicRank	27.6	31.5	27.9	14.9	10.3	12.1	35.0	37.5	35.6	11.7	21.7	15.1

Backups

Contributions Evaluation

Method	Inspec			SemEval			WikiNews			DEFT		
	P	R	F	P	R	F	P	R	F	P	R	F
SingleRank	34.8	40.4	35.2	4.6	3.2	3.7	19.4	20.7	19.7	4.5	9.0	5.9
+phrases	21.5	25.9	22.1	9.6	7.0	8.0	28.6	30.1	28.9	10.5	19.7	13.5
+topics	26.6	30.2	26.8	14.7	10.2	11.9	31.0	32.8	31.4	11.5	21.4	14.8
+complete	34.9	41.0	35.5	5.5	3.8	4.4	20.0	21.4	20.3	4.4	9.0	5.8
TopicRank	27.6	31.5	27.9	14.9	10.3	12.1	35.0	37.5	35.6	11.7	21.7	15.1

Backups

Keyphrase Selection Evaluation

Keyphrase selection	Inspec			SemEval			WikiNews			DEFT		
	P	R	F	P	R	F	P	R	F	P	R	F
First position	27.6	31.5	27.9	14.9	10.3	12.1	35.0	37.5	35.6	11.7	21.7	15.1
Frequency	26.7	30.2	26.8	1.7	1.2	1.4	25.7	27.6	26.2	1.9	3.8	2.5
Centroid	24.5	28.0	24.7	1.9	1.2	1.5	28.1	29.9	28.5	2.6	5.0	3.4
Upper bound	36.4	39.0	35.6	37.6	25.8	30.3	42.5	44.8	42.9	14.9	28.0	19.3

References

Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. Clustering to Find Exemplar Terms for Keyphrase Extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1, pages 257–266, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://dl.acm.org/citation.cfm?id=1699510.1699

References

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. Automatic Keyphrase Extraction Via Topic Decomposition. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 366–376, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1870658.1870

Rada Mihalcea and Paul Tarau. TextRank: Bringing Order Into Texts. In Dekang Lin and Dekai Wu, editors, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

References

Takashi Tomokiyo and Matthew Hurst. A Language Model Approach to Keyphrase Extraction. In Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18, pages 33–40, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/1119282.1119287.

George Tsatsaronis, Iraklis Varlamis, and Kjetil Nørvåg. SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs. In Proceedings of the 23rd International Conference on Computational Linguistics, pages 1074–1082, Stroudsburg, PA, USA, 2010.

References

Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1873781.1873

Xiaojun Wan and Jianguo Xiao. Single Document Keyphrase Extraction Using Neighborhood Knowledge. In Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2, pages 855–860. AAAI Press, 2008. ISBN 978-1-57735-368-3. URL http://dl.acm.org/citation.cfm?id=1620163.1620