Mining Quality Phrases from Massive Text Corpora

Jialu Liu*, Jingbo Shang*, Chi Wang, Xiang Ren, Jiawei Han

University of Illinois at Urbana-Champaign

SIGMOD 2015, May 2015

* Equal Contribution

Outline

- Motivation: Why Phrase Mining?
- SegPhrase+: Methodology
- Performance Study and Experimental Results
- Discussion and Future Work

Why Phrase Mining?

- Unigrams vs. phrases
  - **Unigrams** (single words) are *ambiguous*
    - Example: “United”: United States? United Airline? United Parcel Service?
  - **Phrase**: A natural, meaningful, *unambiguous* semantic unit
    - Example: “United States” vs. “United Airline”
- Mining semantically meaningful phrases
  - Transform text data from *word granularity* to *phrase granularity*
  - Enhance the power and efficiency at manipulating unstructured data using database technology

Mining Phrases: Why Not Use NLP Methods?

- Phrase mining was originated from the NLP community
  - Name Entity Recognition (NER) can only identify noun phrases
  - Chunking can provide some phrase candidates
- Most NLP methods need heavy training and complex labeling
  - Costly and may not be transferable
  - May not fit domain-specific, dynamic, emerging applications
    - Scientific domains
    - Query logs
    - Social media, e.g., Yelp, Twitter

Mining Phrases: Why Not Use Raw Frequency Based Methods?

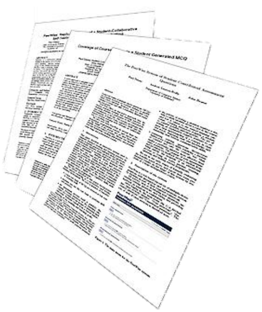
- Traditional data-driven approaches
  - Frequent pattern mining
    - If AB is frequent, likely AB could be a phrase
- Raw frequency could NOT reflect the quality of phrases
  - E.g., freq(vector machine) ≥ freq(support vector machine)
  - Need to rectify the frequency based on segmentation results
- Phrasal segmentation will tell
  - Some words should be treated as a whole phrase whereas others are still unigrams

Outline

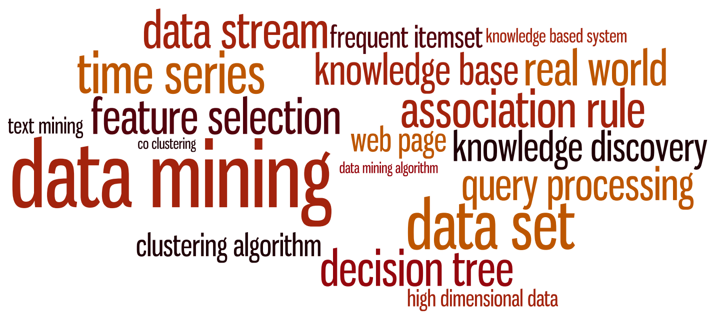
- Motivation: Why Phrase Mining?
- SegPhrase+: Methodology
- Performance Study and Experimental Results
- Discussion and Future Work

SegPhrase: From Raw Corpus to Quality Phrases and Segmented Corpus

Raw Corpus



Quality Phrases



Segmented Corpus

Document 1
Citation recommendation is an interesting but challenging research problem in data mining area.

Document 2
In this study, we investigate the problem in the context of heterogeneous information networks using data mining technique.

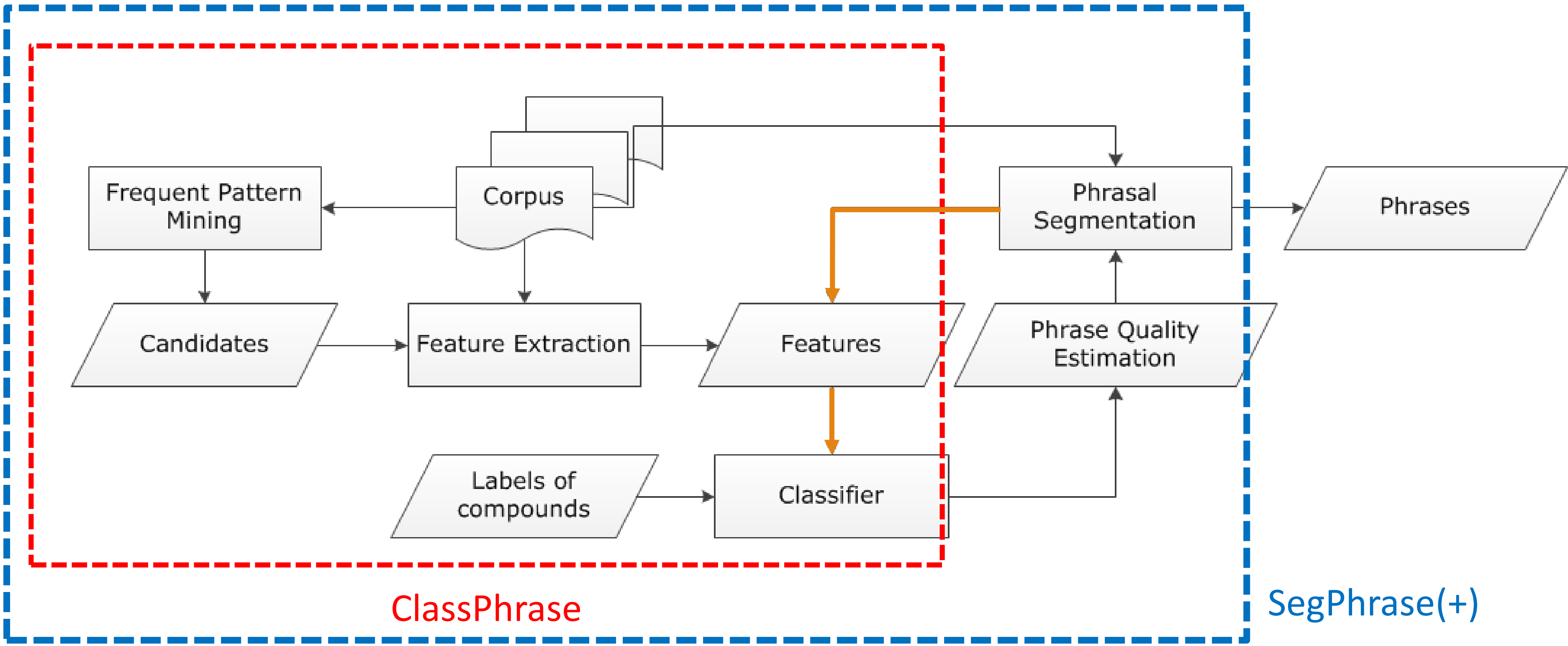
Document 3
Principal Component Analysis is a linear dimensionality reduction technique commonly used in machine learning applications.

Input Raw Corpus → **Quality Phrases** ↔ **Segmented Corpus**

Phrase Mining **Phrasal Segmentation**

SegPhrase: The Overall Framework

- ClassPhrase: Frequent pattern mining, feature extraction, classification
- SegPhrase: Phrasal segmentation and phrase quality estimation
- SegPhrase+: One more round to enhance mined phrase quality



What Kind of Phrases Are of "High Quality"?

- Judging the quality of phrases
  - **Popularity**
    - "information retrieval" vs. "cross-language information retrieval"
  - **Concordance**
    - "powerful tea" vs. "strong tea"
    - "active learning" vs. "learning classification"
  - **Informativeness**
    - "this paper" (frequent but not discriminative, not informative)
  - **Completeness**
    - "vector machine" vs. "support vector machine"

ClassPhrase I: Pattern Mining for Candidate Set

- Build a candidate phrases set by frequent pattern mining
  - Mining frequent *k*-grams
    - *k* is typically small, e.g. 6 in our experiments
- **Popularity** measured by *raw* frequent words and phrases mined from the corpus

ClassPhrase II: Feature Extraction: Concordance

- Partition a phrase into two parts to check whether the co-occurrence is significantly higher than pure random
  - $\boxed{\text{support vector}}_{u_l}\ \boxed{\text{machine}}_{u_r}$ $\boxed{\text{this paper}}_{u_l}\ \boxed{\text{demonstrates}}_{u_r}$

$$\langle u_l, u_r \rangle = \arg \min_{u_l \oplus u_r = v} \log \frac{p(v)}{p(u_l)p(u_r)}$$

  - Pointwise mutual information:

$$PMI(u_l, u_r) = \log \frac{p(v)}{p(u_l)p(u_r)}$$

  - Pointwise KL divergence:

$$PKL(v \| \langle u_l, u_r \rangle) = p(v) \log \frac{p(v)}{p(u_l)p(u_r)}$$

- The additional *p*(*v*) is multiplied with pointwise mutual information, leading to less bias towards rare-occurred phrases

ClassPhrase II: Feature Extraction: Informativeness

- Deriving Informativeness
  - Quality phrases typically start and end with a non-stopword
    - “machine learning is” v.s. “machine learning”
  - Use average IDF over words in the phrase to measure the semantics
  - Usually, the probabilities of a quality phrase in quotes, brackets, or connected by dash should be higher (punctuations information)
    - “state-of-the-art”
- We can also incorporate features using some NLP techniques, such as POS tagging, chunking, and semantic parsing

ClassPhrase III: Classifier

- **Limited Training**
  - Labels: Whether a phrase is a quality one or not
    - "support vector machine": 1
    - "the experiment shows": 0
  - For ~1GB corpus, only 300 labels
- Random Forest as our classifier
  - Predicted phrase quality scores lie in [0, 1]
  - Bootstrap many different datasets from limited labels

SegPhrase: Why Do We Need Phrasal Segmentation in Corpus?

- Phrasal segmentation can tell which phrase is more appropriate
  - Ex: A standard [feature vector] [machine learning] setup is used to describe...

Not counted towards the rectified frequency

- Rectified phrase frequency (expected influence)
  - Example:

sequence	frequency	phrase?	rectified
support vector machine	100	yes	80
support vector	160	yes	50
vector machine	150	no	6
support	500	N/A	150
vector	1000	N/A	200
machine	1000	N/A	150

SegPhrase: Segmentation of Phrases

- Partition a sequence of words by maximizing the likelihood
  - Considering
    - Phrase quality score
      - ClassPhrase assigns a **quality score** for each phrase
    - Probability in corpus
    - Length penalty
      - **length penalty** α: when $\alpha > 1$, it favors shorter phrases
- Filter out phrases with low rectified frequency
  - Bad phrases are expected to rarely occur in the segmentation results

SegPhrase+: Enhancing Phrasal Segmentation

- SegPhrase+: One more round for enhanced phrasal segmentation
- **Feedback**
  - Using rectified frequency, re-compute those features previously computing based on raw frequency
- Process
  - Classification → Phrasal segmentation // SegPhrase
    → Classification → Phrasal segmentation // SegPhrase+
- **Effects** on computing quality scores
  - np hard in the strong sense ↗
  - ~~np hard in the strong~~ ↘
  - data base management system ↗

Outline

- Motivation: Why Phrase Mining?
- SegPhrase+: Methodology
- Performance Study and Experimental Results
- Discussion and Future Work

Performance Study: Methods to Be Compared

- Other phase mining methods: Methods to be compared
  - NLP chunking based methods
    - Chunks as candidates
    - Sorted by **TF-IDF** and **C-value** (K. Frantzi et al., 2000)
  - Unsupervised raw frequency based methods
    - **ConExtr** (A. Parameswaran et al., VLDB 2010)
    - **ToPMine** (A. El-Kishky et al., VLDB 2015)
  - Supervised method
    - **KEA**, designed for single document keyphrases (O. Medelyan & I. H. Witten, 2006)

Performance Study: Experimental Setting

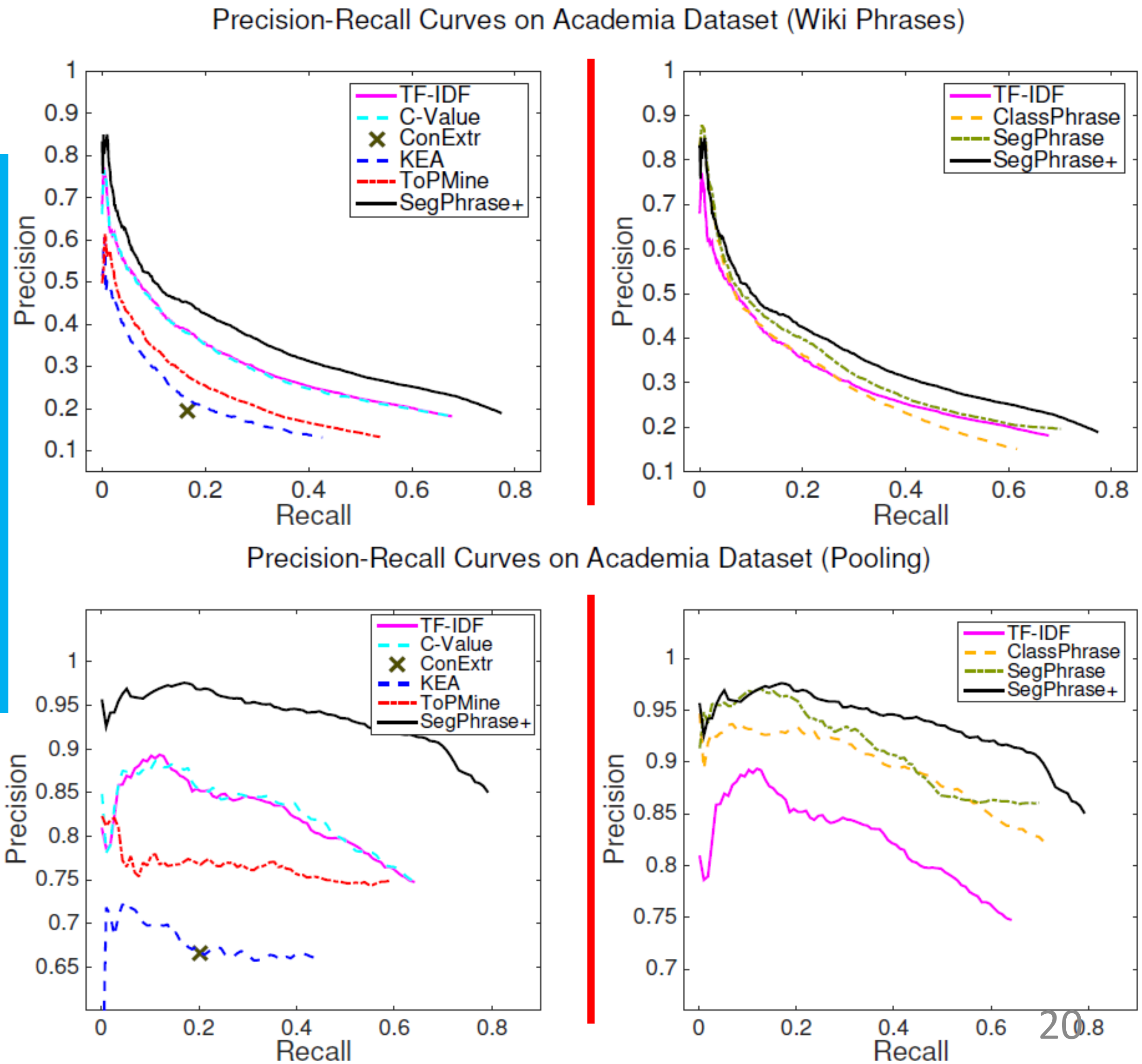
- Datasets

Dataset	#docs	#words	#labels
DBLP	2.77M	91.6M	300
Yelp	4.75M	145.1M	300

- Popular Wiki Phrases
  - Based on internal links
  - ~7K high quality phrases
- Pooling
  - Sampled 500 * 7 **Wiki-uncovered** phrases
  - Evaluated by 3 reviewers independently

Performance: Precision Recall Curves on DBLP

Compare with other baselines
TF-IDF
C-Value
ConExtr
KEA
ToPMine
SegPhrase+

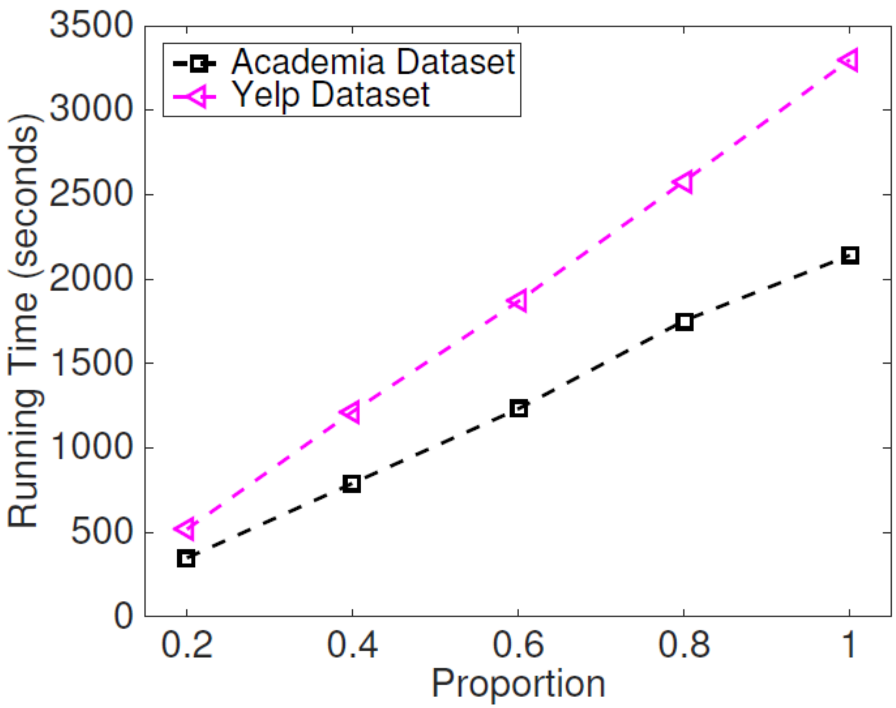


Compare with our 3 variations
TF-IDF
ClassPhrase
SegPhrase
SegPhrase+

20

Performance Study: Processing Efficiency

- SegPhrase+ is linear to the size of corpus!



dataset	file size	#words	time
Academia	613MB	91.6M	0.595h
Yelp	750MB	145.1M	0.917h
Wikipedia	20.23GB	3.26G	28.08h

Experimental Results: Interesting Phrases Generated (From the Titles and Abstracts of SIGMOD)

Query	SIGMOD	
Method	SegPhrase+	Chunking (TF-IDF & C-Value)
1	data base	data base
2	database system	database system
3	relational database	query processing
4	query optimization	query optimization
5	query processing	relational database
…	…	…
51	sql server	database technology
52	**relational data**	database server
53	data structure	**large volume**
54	**join query**	performance study
55	web service	web service
…	…	…
201	**high dimensional data**	**efficient implementation**
202	location based service	sensor network
203	xml schema	**large collection**
204	**two phase locking**	**important issue**
205	deep web	frequent itemset
…	…	…

Only in SegPhrase+

Only in Chunking

Experimental Results: Interesting Phrases Generated (From the Titles and Abstracts of SIGKDD)

Query	SIGKDD	
Method	SegPhrase+	Chunking (TF-IDF & C-Value)
1	data mining	data mining
2	data set	association rule
3	association rule	knowledge discovery
4	knowledge discovery	frequent itemset
5	**time series**	decision tree
…	…	…
51	association rule mining	search space
52	rule set	domain knowledge
53	concept drift	**importnant problem**
54	knowledge acquisition	concurrency control
55	**gene expression data**	conceptual graph
…	…	…
201	web content	optimal solution
202	**frequent subgraph**	semantic relationship
203	intrusion detection	**effective way**
204	**categorical attribute**	space complexity
205	user preference	**small set**
…	…	…

Only in SegPhrase+

Only in Chunking

23

Experimental Results: Similarity Search

- Find high-quality similar phrases based on user's phrase query
  - In response to a user's phrase query, SegPhrase+ generates high quality, semantically similar phrases
  - In DBLP, query on "data mining" and "OLAP"
  - In Yelp, query on "blu-ray", "noodle", and "valet parking"

Query	data mining		olap	
Method	SegPhrase+	Chunking	SegPhrase+	Chunking
1	knowledge discovery	driven methodologies	data warehouse	warehouses
2	text mining	text mining	online analytical processing	clustcube
3	web mining	financial investment	data cube	rolap
4	machine learning	knowledge discovery	olap queries	online analytical processing
5	data mining techniques	building knowledge	multidimensional databases	analytical processing

Query	blu-ray		noodle		valet parking	
Method	SegPhrase+	Chunking	SegPhrase+	Chunking	SegPhrase+	Chunking
1	dvd	new microwave	ramen	noodle soup	valet	huge lot
2	vhs	lifetime warranty	noodle soup	asian noodle	self-parking	private lot
3	cd	recliner	rice noodle	beef noodle	valet service	self-parking
4	new release	battery	egg noodle	stir fry	free valet parking	valet
5	sony	new battery	pasta	fish ball	covered parking	front lot

Outline

- Motivation: Why Phrase Mining?
- SegPhrase+: Methodology
- Performance Study and Experimental Results
- Discussion and Future Work

Recent Progress after SIGMOD Final Version

- Distant Training: No need of human labeling
  - Training using general knowledge bases
    - E.g., Freebase, Wikipedia
- Quality Estimation for Unigrams
  - Integration of phrases and unigrams in one uniform framework
- Multi-languages: Beyond English corpus
  - Extensible to mining quality phrases in multiple languages
  - Recent progress: SegPhrase+ works on Chinese and Arabic

Experimental Results: High Quality Phrases Generated (From Chinese Wikipedia)

Rank	Phrase	In English
…	…	…
62	首席_执行官	CEO
63	中间_偏右	Middle-right
…	…	…
84	百度_百科	Baidu Pedia
85	热带_气旋	Tropical cyclone
86	中国科学院_院士	Fellow of Chinese Academy of Sciences
…	…	…
1001	十大_中文_金曲	Top-10 Chinese Songs
1002	全球_资讯网	Global News Website
1003	天一阁_藏_明代_科举_录_选刊	A Chinese book name
…	…	…
9934	国家_戏剧_院	National Theater
9935	谢谢_你	Thank you
…	…	…

Conclusions and Future Work

- SegPhrase+: A new phrase mining framework
  - Integrating phrase mining with phrasal segmentation
  - Requires only limited training or distant training
  - Generates high-quality phrases, close to human judgement
  - Linearly scalable on time and space
- Looking forward: High-quality, scalable phrase mining
  - Facilitate entity recognition and typing in large corpora
  - Transform massive unstructured data into semi-structured knowledge networks

References

- A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han. Scalable topical phrase mining from text corpora. *VLDB*, 8(3), Aug. 2015
- A. Parameswaran, H. Garcia-Molina, and A. Rajaraman. Towards the web of concepts: Extracting concepts from large datasets. *VLDB*, 3(1-2), Sept. 2010
- Medelyan, O., & Witten, I. H. (2006) Thesaurus based automatic keyphrase indexing. In Proc. of the 6th ACM/IEEE-CS Joint Conf. on Digital Libraries (pp. 296-297)
- Frantzi, K., Ananiadou, S., & Mima, H. (2000) Automatic recognition of multi-word terms: the c-value/nc-value method. Int. Journal on Digital Libraries, 3(2), 115-130