Feature Associations in Graph Structures for Unsupervised Entity Disambiguation

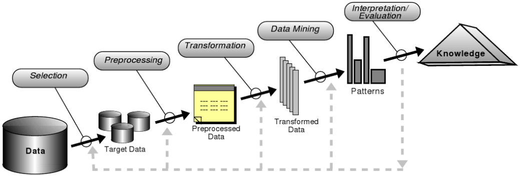
Roman Kern
rkern@know-center.at

WIQE10 / 2010-09-14

Overview

Anatomy of a Knowledge Discovery Application

- Input: Data stored in repositories
  - Structured vs. unstructured data
  - Textual vs. multi-media content
  - Single vs. multiple repositories
- Preprocessing of input into data-structures suitable for algorithms
- Apply algorithms on data-structures
- Output: Visualize & store result



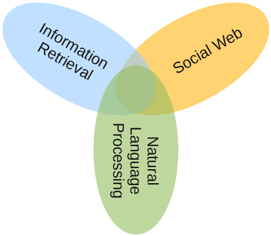
[Fayyad et al. 96]

Feature Engineering

- Transform data into features
  - Feature extraction: Words, Syntax, Statistics, ...
  - Feature representation: Plain Text, Arrays, Matrix, Graph, ...
- Specific algorithms need specific data-structures
  - High Level: Information Extraction, Classification, Clustering, Information Retrieval, ...
  - Low Level: *SVD*, *EVD*, *SVM*, *HAC*, *BM*25, *LSA*, *LDA*, *TFIDF*, *CRF*, *KNN*, *HMM*, ...

- *Example:* Vector Space Model
  - Input: Documents
  - Features: Terms
  - Data-structure: Matrix

Feature Associations



Relationships between features

- Additional transformation step
- Network of features
- *Example:* Term co-occurrences

Goal: Framework for feature associations

- Calculate feature associations
- Provide data-structure for feature associations
- Support feature engineering
  - Feature analysis
  - Feature synthesis
- Support application development
  - Common data-structure for algorithms from various domains

Algorithmic Issues

How to represent different features?

- Associations between features of different types
- More features could lead to better results
  - *Example:* String kernels for classification
- But: More features definitely lead to more expensive computation

How to integrate external knowledge?

- WordNet, ConceptNet, Linked Data, LDAP, ...

How to calculate the association weight?

- Correlation, statistical tests, probabilities, ...

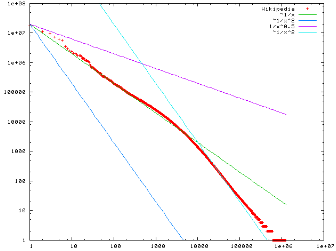
Practical Issues

How to deal with data that does not fit into main memory?

- Enterprise scale
- *Example:* English Wikipedia: 10^7 documents, 10^6 terms

How to integrate (exploit) heuristics?

- Strong (naive) independence assumption, Zipf's law, Heaps' law, small world networks, distributional hypothesis, ...



Approach - Overview

Feature association framework

- Calculate feature associations
  - Input: Extracted features in graph-like structures
  - Output: Feature network
- Access feature associations
  - Traverse feature network

Solves algorithmic and practical issues

- Provides an scalable algorithmic approach for large scale datasets
- Flexible to allow the integration of rich set of features and external sources
- Allows the integration of a range of graph operations to build the association network

Approach - Generalizations

Starting Point

- Vector Space Model: Inverted Index
- Simple feature representation: $Matrix_{Documents \times Terms}$
- Simple feature operations: $cos_{sim}(row(M, i), row(M, j))$

Generalize Feature Representation

- Generalization of the simple matrix model
- Allows integration of additional information, e.g. term positions, external sources, linguistic annotations, ...

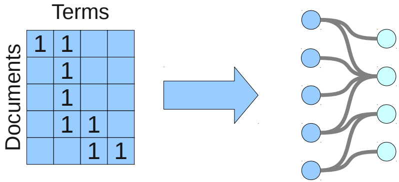
Generalize Feature Operations

- Generalization of the operations on the features
- Allows integration of algorithms, e.g. Levenshtein edit distance, SVD, clustering, ...

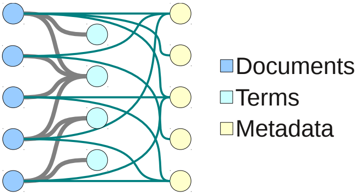
Generalize Feature Representation

Feature Data-Structure

- Matrix can be transformed into a bi-partite graph



- Bi-partite graph can be generalized to a n-partite graph



Generalize Feature Operations

Matrix Feature Function

- Matrix multiplication

$$M_{n,n} = M_{m,n} \times M_{m,n}^T$$

- General matrix transformation

$$M_{n',m'} = f(M_{m,n}, M_{m,n}^T)$$

- Simple Example - Matrix transposition

$$M_{n,m} = M_{m,n}^T$$

Not only for matrices, but graphs too.

Feature Operation Functions

- Feature association function - $f(i,j)$

$$f(node_i, node_j) = w_{global}(a(node_i, node_j), \mathcal{G})$$

$$a(node_i, node_j) = w_{aggregate}(\{w_{combine}(l(i), l(j))\})$$

$$l(node_x) = w_{local}(node_x, \mathcal{L})$$

- Input variables
  - Local - $\mathcal{L}$: word-forms, position, term frequency, document length, ...
  - Global - $\mathcal{G}$: document frequency, dispersion, co-occurrence count, average document length, ...
- Examples
  - Cosine Similarity, Jaccard, Windowed Co-Occurrence, Poisson, Pascal, Binomial, PMI, Conditional Probability, Conditional Entropy, Mutual Information, χ^2, Log Odds, ...

Properties

Runtime Complexity

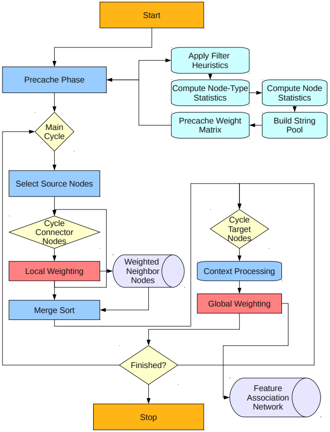
- Runtime complexity of $\mathcal{O}(n^2 * m)$
- Wikipedia: 10^{19} Operations

Algorithm

- Number of heuristics to keep computation feasible
  - Expects power law
  - Expects globally sparse, but locally highly connected
- Execution can be done in parallel
- Map-Reduce friendly

Implementation

- Implementation tailored towards contemporary computer architecture: Memory Access $\ll$ Disk Access
- CPU & IO bound



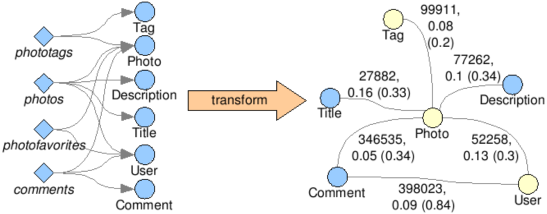
Tag Recommender - Tagr (SASU)

- Item based tag recommender system
  - Input data: Folksonomy (Flickr subset)
  - Features: Tags, Photos
  - Data-Structure: Bipartite Graph (Matrix)
  - Feature function: $w_{i,j} = \frac{sharedPhotos_{i,j}}{mean(photoCount_i, photoCount_j)}$
  - Feature association retrieval: Lookup

Recommending tags for pictures based on text, visual content and user context.
Lindstaedt, Pammer, Moerzinger, Kern, Mülner, and Wagner [2008]

Folksonomy Analysis

- Statistical Analysis of a Folksonomy
  - Input data: Folksonomy (Flickr subset) stored in SQL-database
  - Features: Tags, Photos, Users, Title, Description and Comments
  - Data-Structure: N-Bipartite Graph
  - Feature function: Cosine
  - Feature association retrieval: Spreading Activation, Distance

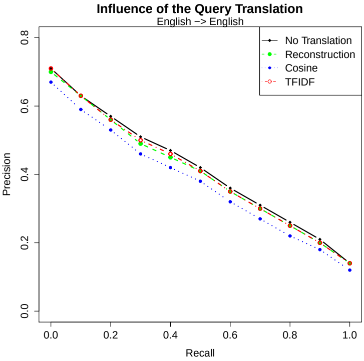


Extending Folksonomies for Image Tagging.
Kern, Granitzer, and Pammer [2008]

Machine Translation

- Word alignment for query translation

- Input data: multilingual corpus (Wikipedia, Europarl)
- Features: English words, Spanish words
- Data-Structure: article aligned corpus, 3-partite graph
- Feature function: Cosine, Correlation, TFIDF
- Feature association retrieval: Spreading Activation



Crosslanguage Retrieval based on Wikipedia Statistics. Juffinger, Kern, and Granitzer [2008a]
Exploiting Cooccurrence on Corpus and Document Level for Fair Crosslanguage Retrieval. Juffinger, Kern, and Granitzer [2008b]

Information Retrieval

- Global query expansion for cross-lingual information retrieval
  - Textual corpus (Glasgow Harald, LA Times)
  - English words & positions, PMI
  - Monolingual Performance

Query Expansion	MAP	GMAP	Wilcoxon	Randomized
Baseline	0.4022	0.1805	-	-
WSD WordNet	0.4070	0.1869	0.0119	0.0147
Co-occurrence Terms	0.4170	0.1864	0.0001	0.0196

  - Crosslingual Performance

Query Expansion	MAP	GMAP	Wilcoxon	Randomized
Baseline	0.2885	0.0762	-	-
WSD WordNet	0.2933	0.0773	0.2187	0.0056
Co-occurrence Terms	0.2917	0.0718	0.0090	0.0252

Application of Axiomatic Approaches to Crosslanguage Retrieval.
Kern, Juffinger, and Granitzer [2009a]
Evaluation of Axiomatic Approaches to Crosslanguage Retrieval.
Kern, Juffinger, and Granitzer [2009b]

Crosslingual Plagiarism Detection

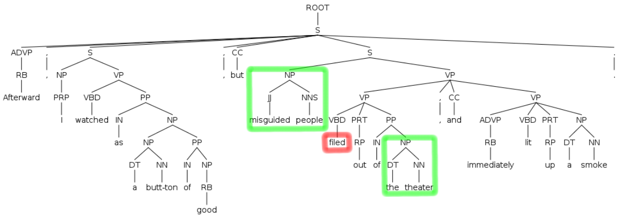
- Goal: Retrive word translation candidates to detect crosslingual plagiarism
- Features: word alignment candidates
- Data-Structure: sentence aligned corpus (Europarl)
- Feature Function: HMM based word alignment algorithm (BerkeleyAligner)

External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation System
Muhr, Kern, Zechner, and Granitzer [2010]

Unsupervised Entity Disambiguation - 1/2

Word sense induction and discrimination

- Goal: Identify the individual senses of an ambiguous word and label unseen instance with one of them
- Features: Grammatical dependencies, (expanded) sentence phrase terms



relation	gov ▲	dep
pobj	as-5	butt-ton-7
det	butt-ton-7	a-6
prep	butt-ton-7	of-8
nsubj	filed-14	people-13
prt	filed-14	out-15
prep	filed-14	of-16
cc	filed-14	and-20
conj	filed-14	lit-22
advmod	lit-22	immediat...
prt	lit-22	up-23
dobj	lit-22	smoke-25
pobj	of-16	theater-18
pobj	of-8	good-9
amod	people-13	misguide...
det	smoke-25	a-24
det	theater-18	the-17
advmod	watched-4	Afterwar...
nsubj	watched-4	I-3
prep	watched-4	as-5
cc	watched-4	but-11
conj	watched-4	filed-14

Unsupervised Entity Disambiguation - 2/2

Word sense induction and discrimination

- Sense induction:
  - Extract local sub-graphs
  - Cluster sub-graphs
  - Generate new features out of existing features
  - Senses are additional features in the feature association network
- Sense discrimination:
  - Distance based similarity search

KCDC: Word Sense Induction by Using Grammatical Dependencies and Sentence Phrase Structure. Kern, Muhr, and Granitzer [2010]

Results

Classification

Domain	Application	Key Results	Benefits
Social Web	Tag Recommender	Proof of concept	Easy exchange of similarity function
Social Web	Folksonomy Analysis	Deeper understanding of folksonomies, Base for recommender systems	Integration of additional features
Information Retrieval	Query Translation	Good translation performance	Simple mapping of aligned documents, Efficient lookup
Information Retrieval	Query Expansion	Improved Performance over baseline	Integration of task-specific weighting function
Natural Language Processing	Crosslingual Plagiarism Detection	Lookup runtime performance	Integration into real-world systems
Natural Language Processing	Word Sense Induction and Discrimination	State-of-the-art performance	Integration of different features, Integration of different algorithms

Conclusions

- Algorithmic approach
  - Calculate feature associations
  - Traverse feature association networks
- Goals
  - Scalable
  - Flexible
  - Usable
- Applications
  - Different domains related to knowledge discovery
  - Real-world benefit

The End

Thank you!

References

A. Juffinger, R. Kern, and M. Granitzer. Crosslanguage Retrieval based on Wikipedia Statistics. In *Proc. of 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, 17-19 September, Aarhus, Denmark*, 2008a.

A. Juffinger, R. Kern, and M. Granitzer. Exploiting Cooccurrence on Corpus and Document Level for Fair Crosslanguage Retrieval. In *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark*, 2008b.

R. Kern, M. Granitzer, and V. Pammer. Extending Folksonomies for Image Tagging. In *WIAMIS 2008 , Special Session on Multimedia Metadata Management & Retrieval*. IEEE Computer Society, 2008.

R. Kern, A. Juffinger, and M. Granitzer. Evaluation of Axiomatic Approaches to Crosslanguage Retrieval. In *Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments*, 2009a.

R. Kern, A. Juffinger, and M. Granitzer. Application of Axiomatic Approaches to Crosslanguage Retrieval. In *Working Notes for the CLEF 2009 Workshop*, 2009b.

R. Kern, M. Muhr, and M. Granitzer. KCDC: Word Sense Induction by Using Grammatical Dependencies and Sentence Phrase Structure. In *Proceedings of SemEval-2*, Uppsala, Sweden, ACL, 2010.

S. N. Lindstaedt, V. Pammer, R. Moerzinger, R. Kern, H. Mülner, and C. Wagner. Recommending tags for pictures based on text, visual content and user context. In *Proceedings of the Third International Conference on Internet and Web Applications and Services (ICIW 2008)*, pages 506—-511. IEEE Computer Society Press, 2008.

M. Muhr, R. Kern, M. Zechner, and M. Granitzer. External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation System. In *Lab Report for PAN at CLEF 2010*, 2010.