Content-Based Multimedia Indexing (CBMI)

On Using Statistical Semantic on Domain-Specific Information Retrieval

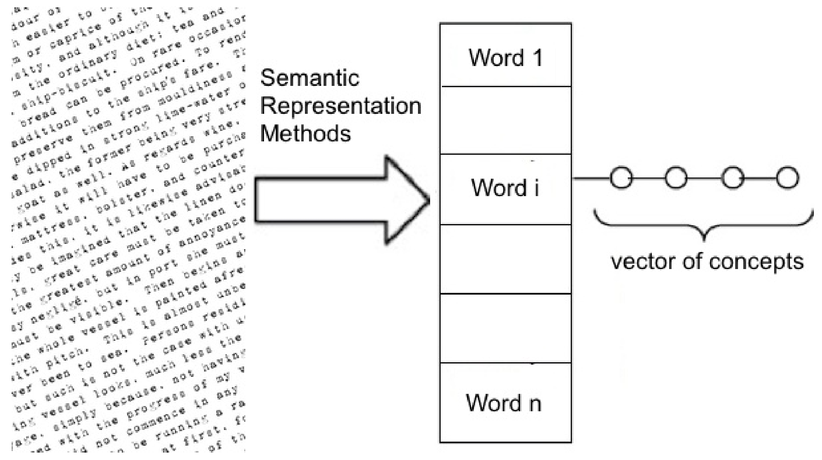
Vienna University of Technology
Institute of Software Technology & Interactive System
Information & Software Engineering Group

Navid Rekabsaz, Ralf Bierig, Bogdan Ionescu, Allan Hanbury and Mihai Lupu

Background & Motivation

Statistical Semantic Representation

- Mainstream in IR: term-frequence-based
- Semantic Repres.: representing text elements as a vector of concepts
- Stat. Sem. Repres. defines meanings based on the occurance of words in contexts:
  - Word2Vec(W2V): state-of-the-art, deep learning
  - Random Indexing(RI): refines random vectors



Motivation

- Social Image Retrieval: different words in the text refer to the same concept
- Exploiting the benefits of Word-to-Word semantic similarity to Document-to-Document semantical similarity

Research Objectives

- Does stat. sem. outperform term-freq.-based?
- Which of W2V/RI performs better?

Document-to-Document Semantic Similarity

- SimAgg

$$V_A = \sum_{t \in A} idf(t) * V_t$$

$$SimAgg(A, B) = Cos(V_A, V_B)$$

- SimGreedy

$$SimGreedy(A, B) = \frac{\sum_{t \in A} idf(t) * maxSim(t, B)}{\sum_{t \in A} idf(t)}$$

$$SimGreedy = \frac{SimGreedy(A, B) + SimGreedy(B, A)}{2}$$

Experimental Results

Sanity Check

- SemEval 2014 Multilingual Semantic Textual Similarity-Task 10
- Word2Vec and RandomIndexing trained on Wikipedia corpus
- SimGreedy+W2V: ranked 11th between SemEval 38 runs, best run without NLP & knowledge-bases

Representation	Dim	SimAgg	SimGreedy
RI	600	0.691	0.706
RI	200	0.678	0.702
W2V	600	0.685	**0.715**
W2V	200	0.654	0.715

Social Image Retrieval

- MediaEval Retrieving Diverse Social Images Task 2013/14
- Word2Vec & Random Indexing Trained on Wikipedia corpus
- Best practice on Solr as baseline
- SimGreedy in combination with Word2Vec outperforms the baseline

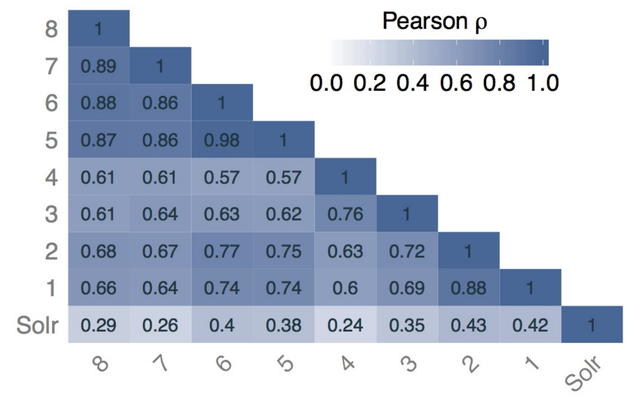
Repr.	Dim	SimAgg	SimGreedy	(Q,D)	(D,Q)
RI	200	0.774 (1)	†0.788 (5)	0.704	0.766
RI	600	0.766 (2)	†0.787 (6)	0.703	0.769
W2V	200	0.778 (3)	†**0.795** (7)	0.690	0.760
W2V	600	0.779 (4)	†0.793 (8)	0.693	0.757

- Repeat the experiment only on 2014 dataset
- Word2Vec & Random Indexing Trained on Wikipedia & MediaEval corpora
- Ranked 1st between 41 runs, even between the ones with image features

Corpus	Repres.	Dim	SimAgg	SimGreedy
Wiki	RI	200	0.795	0.833
Wiki	W2V	200	0.788	0.813
MediaEval	RI	200	0.840	0.820
MediaEval	W2V	200	0.831	**0.848**

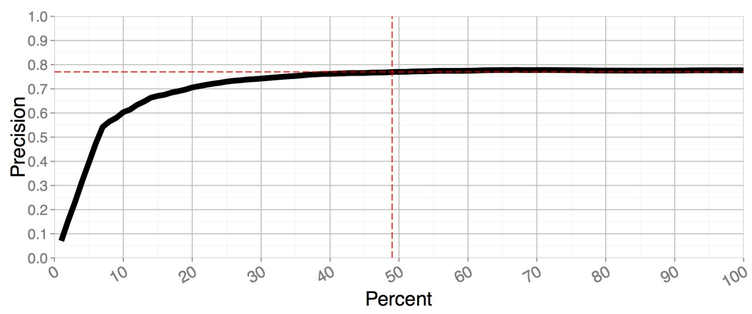
Results Correlations

- The results of SimGreedy are the most similar regardless of Repres. method or dimensions



Optimization Methods

- Combining SimAgg with SimGreedy to exploit the benefits of both
- Finding the cut-off point between the algorithms



Conclusion

- Exploring the effect of vector representation of words in sematnic social image retrieval
- SimGreedy significantly better than SimAgg & Solr
- Weak effect of dimentionality & semantic representation methods

Contact: rekabsaz@ifs.tuwien.ac.at