WikiWalk: Random walks on Wikipedia for Semantic Relatedness

HS Graph-based Methods for Natural Language Processing

Katharina Wäschle

Seminar für Computerlinguistik

11/4/2010

Outline

Outline

Motivation

Motivation many NLP tasks use numerical measures for semantic relatedness, e.g. *text summarization*, *information retrieval* or *word sense disambiguation*

Motivation

Motivation many NLP tasks use numerical measures for semantic relatedness, e.g. *text summarization*, *information retrieval* or *word sense disambiguation*

Idea capture world knowldege encoded in the Wikipedia link structure

Motivation

Motivation many NLP tasks use numerical measures for semantic relatedness, e.g. *text summarization*, *information retrieval* or *word sense disambiguation*

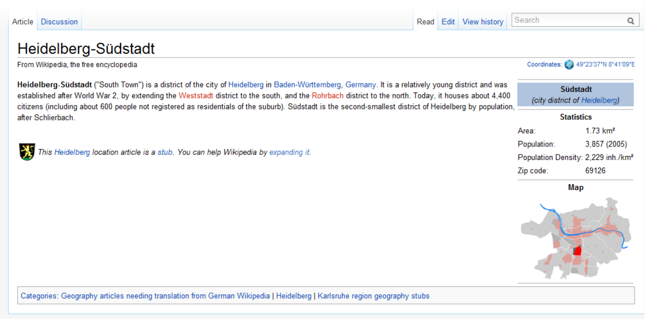
Idea capture world knowldege encoded in the Wikipedia link structure

Approach compare *Personalized PageRank* vectors of random walks on a graph derived from Wikipedi

Outline

Wikipedia

- largest online encyclopedia with articles on a wide variety of topics
- high number of hyperlinks between articles



PageRank

PageRank

- assigns a numerical weighting to each element of a graph in the form of a probability distribution over the nodes, the *PageRank* vector:
  $Pr =$(0.05 0.01 0.0 ... 0.2 ... 0.01 0.05 0.03)

PageRank

- assigns a numerical weighting to each element of a graph in the form of a probability distribution over the nodes, the *PageRank* vector:
  $Pr =($ 0.05 0.01 0.0 ... 0.2 ... 0.01 0.05 0.03 $)$
- analyze the link structure: whenever a link from node i to node j exists , a vote from node i to node j is produced, and the rank of node j increases; the strength of the vote depends on the rank of node i

PageRank

- assigns a numerical weighting to each element of a graph in the form of a probability distribution over the nodes, the *PageRank* vector:
  $Pr =(\ 0.05\ 0.01\ 0.0\ \ldots\ 0.2\ \ldots\ 0.01\ 0.05\ 0.03\ )$
- analyze the link structure: whenever a link from node i to node j exists , a vote from node i to node j is produced, and the rank of node j increases; the strength of the vote depends on the rank of node i
- can be viewed as a random walk on the graph: the vector entry Pr_i denotes the probability that the walker is at node i at a given point in time

PageRank

Calculation of the *PageRank* vector Pr on a Graph with N nodes:

$$Pr = c \cdot M \cdot Pr + (1 - c) \cdot t$$

where

- c is a damping factor,
- M is the $N \times N$ transition probability matrix that indicates the probability of j being the next node, given we are currently on node i
- t is the $1 \times N$ teleport vector.

PageRank is calculated iteratively by computing the above equation successively until convergence (steady state).

Personalized PageRank

- in traditional *PageRank*, t contains a uniform weight distribution over all N nodes, i.e. every value is $\frac{1}{N}$

Personalized PageRank

- in traditional *PageRank*, t contains a uniform weight distribution over all N nodes, i.e. every value is $\frac{1}{N}$
- in *Personalized PageRank*, t can be non-uniform and carry a bias into the resulting *PageRank* vector, e.g. concentrate all the probability mass on a unique node

Outline

Approach

1. Construct a graph from Wikipedia.

Approach

1. Construct a graph from Wikipedia.
2. For each input text:

    Perform a (biased) random walk over the graph with a custom teleport vector to compute a stationary distribution vector.

Approach

1. Construct a graph from Wikipedia.
2. For each input text:

   Perform a (biased) random walk over the graph with a custom teleport vector to compute a stationary distribution vector.
3. The relatedness score is the cosine similarity between the distributions of two texts.

Building the Wikipedia Graph

The Wikipedia graph G is a tuple (V, E) with:

$$V = \text{Wikipedia articles}$$
$$E = \text{links between articles}$$

Building the Wikipedia Graph

The Wikipedia graph G is a tuple (V, E) with:

$$V = \text{Wikipedia articles}$$
$$E = \text{links between articles}$$

Pruning

Discard articles with < 200 non-stop words and < 5 links

Building the Wikipedia Graph

The Wikipedia graph G is a tuple (V, E) with:

$$V = \text{Wikipedia articles}$$
$$E = \text{links between articles}$$

Pruning

Discard articles with < 200 non-stop words and < 5 links

Parameters

- Link Types
- Generality

Link Types

Categorical	links to category pages that classify the article
Infobox	anchors from the infobox section, which lists defining attributes
Content	remaining anchors from the article text

Generality

An article is said to be more *general* than another when the number of inlinks is larger.

Generality

An article is said to be more *general* than another when the number of inlinks is larger.

Notation

$+k$ for links from a source article s to a more general article t, if

$$\frac{\#inlink(t)}{\#inlink(s)} \geq k$$

respectively $-k$, if

$$\frac{\#inlink(s)}{\#inlink(t)} \geq k$$

else $= k$

Approach

1. Construct a graph from Wikipedia.
2. For each input text:

   Perform a (biased) random walk over the graph with a custom teleport vector to compute a stationary distribution vector.
3. The relatedness score is the cosine similarity between the distributions of two texts.

Constructing the Teleport Vector

Two initialization types:

- Dictionary based initialization
- ESA-based initialization

Dictionary based Initialization

Building the Dictionary

- for a target word, compute the set of articles to which the word refers and distribute the probability mass uniformely over the dictionary entries
- references are article title, redirection pages and disambiguation pages as well as anchor text

Dictionary based Initialization

Building the Dictionary

- for a target word, compute the set of articles to which the word refers and distribute the probability mass uniformely over the dictionary entries
- references are article title, redirection pages and disambiguation pages as well as anchor text

Pruning

- eliminate articles whose title contains a space
- remove dictionary entries of articles that account for $< 1\%(10\%)$ of the occurrences of the target word

Explicit Semantic Analysis (ESA)

Represents the meaning of texts in a high-dimensional space of concepts derived from Wikipedia: input texts are represented as weighted vectors of concepts.

- each dimension in an *ESA* vector corresponds to a Wikipedia article
- for a given text T, the values of the dimensions are the similarity of the text with an article text C_j, subject to TF-IDF weighting: $\sum_{w_i \in T} \mathit{tfidf}(w_i, T) \cdot \mathit{tfidf}(w_i, C_j)$
- the relatedness of two texts is computed as the cosine similarity of their ESA vectors

ESA-based Initialization

- ESA maps query text to a weighted vector over Wikipedia articles
- retain only the scores of the top-n scoring articles
- normalize the resulting vector to obtain a probability distribution

Outline

Evaluation

- dictionary based vs ESA-based
- textual similarity (Lee dataset) and lexical similarity (Miller Charles and WordSim-353 word pair dataset)
- comparison with related systems, one WordNet-based and three Wikipedia-based, WLM, ESA and WikiRelate

Dictionary based Initialization

Dictionary	Graph	MC
all	full	0.369
1%	full	0.610
1%, noent	full	0.565 (0.824)
1%	reduced	0.563
1%	reduced +2	0.530
1%	reduced +4	0.601
1%	reduced +8	0.512
1%	reduced +10	0.491 (0.522)
10%	full	0.604 (0.750)
10%	reduced	0.605 (0.751)
10%	reduced +2	0.491 (0.540)
10%	reduced +4	0.476 (0.519)
10%	reduced +8	0.474 (0.506)
10%	reduced +10	0.430 (0.484)
WordNet		0.90 / 0.89
WLM		0.70
ESA		0.72

Figure: lexical similarity

Dictionary	Graph	WordSim-353
1%	full	0.449
1%, noent	full	0.440 (0.634)
1%	reduced	0.485
WordNet		0.55 / 0.66
WLM		0.69
ESA		0.75
WikiRelate		0.50

Figure: lexical similarity

Dictionary	Graph	(Lee et al., 2005)
1%, noent	Full	0.308
1%	Reduced +4	0.269
ESA		0.72

Figure: textual similarity

ESA-based Initialization

Method	Text Sim
ESA@625	.766
ESA@625+Walk All	0.556
ESA@625+Walk Categories	0.410
ESA@625+Walk Content	0.536
ESA@625+Walk Infobox	0.710

Figure: link types

Method	Text Sim
ESA@625	0.766
ESA@625+Walk Cat@+6	0.770
ESA@625+Walk Cat@+6 Inf@=2	0.772
Bag of words (Lee et al., 2005)	0.1-0.5
LDA (Lee et al., 2005)	0.60
ESA *	0.72

Figure: comparison

Generality of *Category* links

	$+k$	$-k$	$=k$
$k=2$	0.760	0.685	0.462
$k=4$	0.766	0.699	0.356
$k=6$	**0.771**	0.729	0.334
$k=8$	0.768	0.729	0.352
$k=10$	0.768	0.720	0.352

Figure: generality

Outline

Conclusion

- dictionary approach is unable to reach state of the art results on Wikipedia an WordNet
    - evidence, that the article text provides a stronger signal than the link structure
    - the improved results for a pruned dictionary indicate, that only some links are informative
- small improvements over state of the art using ESA vectors as teleport vectors
- future work: finer grained methods of graph construction to improve the value of the Wikipedia link structure

References

- Agirre, E. (2009). Personalized PageRank over WordNet for Similarity and Word Sense Disambiguation. Slides

- Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. The Association for Computer Linguistics, S. 33-41.

- Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In Proceedings of the 20th international joint conference on Artifical intelligence

- Manning, C. Raghavan, P. and Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press. Slides

- Ramage, D., Yeh, E., Agirre, E. and Soroa, A. (2009). WikiWalk: Random walks on Wikipedia for Semantic Relatedness. In Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing