Constructing Concept Clouds from Company Websites

Rosa Tsegaye Aga Chrstian Wartena

[1]Department of Media, Information and Design
Hochschule Hannover

i-KNOW, 2015

Outline

Outline

MOTIVATION

Outline

HARVESTING OF THE WEBSITES

- The company websites are harvested using crawler4j.
- The number of pages to be retrieved 120.
- Breadth-first way is used to get the most important information from the highest levels in the site hierarchy

HARVESTING OF THE WEBSITES

- The company websites are harvested using crawler4j.
- The number of pages to be retrieved 120.
- Breadth-first way is used to get the most important information from the highest levels in the site hierarchy

HARVESTING OF THE WEBSITES

- The company websites are harvested using crawler4j.
- The number of pages to be retrieved 120.
- Breadth-first way is used to get the most important information from the highest levels in the site hierarchy

Outline

CONCEPT EXTRACTION

- STW Thesaurus (for Economics) considered for concept words extraction
- Gate pipeline is used for language identification, sentence splitting, tokenization, lemmatization and annotation
- Concepts in the thesaurus are searched with Apolda.
- Most frequent concepts with their frequency and main STW class are passed for visualization.

CONCEPT EXTRACTION

- STW Thesaurus (for Economics) considered for concept words extraction
- Gate pipeline is used for language identification, sentence splitting, tokenization, lemmatization and annotation
- Concepts in the thesaurus are searched with Apolda.
- Most frequent concepts with their frequency and main STW class are passed for visualization.

CONCEPT EXTRACTION

- STW Thesaurus (for Economics) considered for concept words extraction
- Gate pipeline is used for language identification, sentence splitting, tokenization, lemmatization and annotation
- Concepts in the thesaurus are searched with Apolda.
- Most frequent concepts with their frequency and main STW class are passed for visualization.

CONCEPT EXTRACTION

- STW Thesaurus (for Economics) considered for concept words extraction
- Gate pipeline is used for language identification, sentence splitting, tokenization, lemmatization and annotation
- Concepts in the thesaurus are searched with Apolda.
- Most frequent concepts with their frequency and main STW class are passed for visualization.

Outline

CONTEXT VECTOR CONSTRUCTION AND SIMILARITY COMPUTATION

- Words with frequency range from 4000 to $1 \cdot 10^6$ in the DeWaC Corpus considered as a context feature
- Each word is represented by 16565 features vector.
- The final context feature is constructed on Positive point-wise mutual information (PPMI)

$$ppmi(c, t) = \max\left(\log \frac{p(c|t)}{p(c)}, 0\right). \quad (1)$$

- Similarity distance between the concepts is computed by frequency adjusted cosine.

CONTEXT VECTOR CONSTRUCTION AND SIMILARITY COMPUTATION

- Words with frequency range from 4000 to $1 \cdot 10^6$ in the DeWaC Corpus considered as a context feature
- Each word is represented by 16565 features vector.
- The final context feature is constructed on Positive point-wise mutual information (PPMI)

$$ppmi(c,t) = \max\left(\log \frac{p(c|t)}{p(c)}, 0\right). \quad (1)$$

- Similarity distance between the concepts is computed by frequency adjusted cosine.

CONTEXT VECTOR CONSTRUCTION AND SIMILARITY COMPUTATION

- Words with frequency range from 4000 to $1 \cdot 10^6$ in the DeWaC Corpus considered as a context feature
- Each word is represented by 16565 features vector.
- The final context feature is constructed on Positive point-wise mutual information (PPMI)

$$ppmi(c,t) = \max\left(\log \frac{p(c|t)}{p(c)}, 0\right). \quad (1)$$

- Similarity distance between the concepts is computed by frequency adjusted cosine.

CONTEXT VECTOR CONSTRUCTION AND SIMILARITY COMPUTATION

- Words with frequency range from 4000 to $1 \cdot 10^6$ in the DeWaC Corpus considered as a context feature
- Each word is represented by 16565 features vector.
- The final context feature is constructed on Positive point-wise mutual information (PPMI)

$$ppmi(c,t) = \max\left(\log \frac{p(c|t)}{p(c)}, 0\right). \quad (1)$$

- Similarity distance between the concepts is computed by frequency adjusted cosine.

CONTEXT VECTOR CONSTRUCTION AND SIMILARITY COMPUTATION

- Words with frequency range from 4000 to $1 \cdot 10^6$ in the DeWaC Corpus considered as a context feature
- Each word is represented by 16565 features vector.
- The final context feature is constructed on Positive point-wise mutual information (PPMI)

$$ppmi(c, t) = \max\left(\log \frac{p(c|t)}{p(c)}, 0\right). \quad (1)$$

- Similarity distance between the concepts is computed by frequency adjusted cosine.

CONTEXT VECTOR CONSTRUCTION AND SIMILARITY COMPUTATION

- Words with frequency range from 4000 to $1 \cdot 10^6$ in the DeWaC Corpus considered as a context feature
- Each word is represented by 16565 features vector.
- The final context feature is constructed on Positive point-wise mutual information (PPMI)

$$ppmi(c,t) = \max\left(\log \frac{p(c|t)}{p(c)}, 0\right). \quad (1)$$

- Similarity distance between the concepts is computed by frequency adjusted cosine.

VISUALIZATION

To visualize the concepts:

- ***concept frequency*** *for font size.*
- ***concept category*** *for color*
- ***Semantic similarity*** *for closeness*
- ***Force Atlas algorithm*** *from Gephi, an open source software for graph and network analysis, is used for the layout.*

VISUALIZATION

To visualize the concepts:

- ***concept frequency*** *for font size.*
- ***concept category*** *for color*
- ***Semantic similarity*** *for closeness*
- ***Force Atlas algorithm*** *from Gephi, an open source software for graph and network analysis, is used for the layout.*

VISUALIZATION

To visualize the concepts:

- ***concept frequency*** *for font size.*
- ***concept category*** *for color*
- ***Semantic similarity*** *for closeness*
- ***Force Atlas algorithm*** *from Gephi, an open source software for graph and network analysis, is used for the layout.*

VISUALIZATION

To visualize the concepts:

- ***concept frequency*** *for font size.*
- ***concept category*** *for color*
- ***Semantic similarity*** *for closeness*
- ***Force Atlas algorithm*** *from Gephi, an open source software for graph and network analysis, is used for the layout.*

VISUALIZATION: Wood Pulp Company Website

VISUALIZATION: Orthopedic Shoes Manufacture Company Website

CONCLUSION

1. We introduced **Concept Clouds** with terms representing thesaurus concepts.
2. Concept clouds can represent a company at a glance.
3. Concept clouds for companies can be constructed fully automatically from their web sites.
4. **Distributional semantics** can be used for cloud lay-out.

CONCLUSION

1. We introduced **Concept Clouds** with terms representing thesaurus concepts.
2. Concept clouds can represent a company at a glance.
3. Concept clouds for companies can be constructed fully automatically from their web sites.
4. **Distributional semantics** can be used for cloud lay-out.

CONCLUSION

1. We introduced **Concept Clouds** with terms representing thesaurus concepts.
2. Concept clouds can represent a company at a glance.
3. Concept clouds for companies can be constructed fully automatically from their web sites.
4. **Distributional semantics** can be used for cloud lay-out.

CONCLUSION

1. We introduced **Concept Clouds** with terms representing thesaurus concepts.
2. Concept clouds can represent a company at a glance.
3. Concept clouds for companies can be constructed fully automatically from their web sites.
4. **Distributional semantics** can be used for cloud lay-out.

FUTURE WORK

- Extract more information from the websites: addresses, names of persons, functions, etc.
- Building a focused crawler to find company websites.
- Improve similarity computation.

FUTURE WORK

- Extract more information from the websites: addresses, names of persons, functions, etc.
- Building a focused crawler to find company websites.
- Improve similarity computation.

FUTURE WORK

- Extract more information from the websites: addresses, names of persons, functions, etc.
- Building a focused crawler to find company websites.
- Improve similarity computation.


Thank You!