From ambiguous words to key-concept extraction

Márius Šajgalík, Michal Barla, Mária Bieliková

PeWe@FIIT
personalized web group

SLOVAK UNIVERSITY OF
TECHNOLOGY IN BRATISLAVA
FACULTY OF INFORMATICS
AND INFORMATION TECHNOLOGIES

Our research focus

- **User** modelling
- Natural **language** processing
- "Wild **web**"

- **BrUMo** – browser-based user modelling framework

(Key)words vs. (key-)concepts

- It's easier to extract **keywords** than latent concepts

- **Concepts** are better defined and have higher information content*

* G. Ramakrishnanan and P. Bhattacharyya, "Text representation with wordnet synsets using soft sense disambiguation," in In Proc. of 8th International Conference on Applications of Natural Language to Information Systems (NLDB 2003), 2003, pp. 214–227

From words to concepts

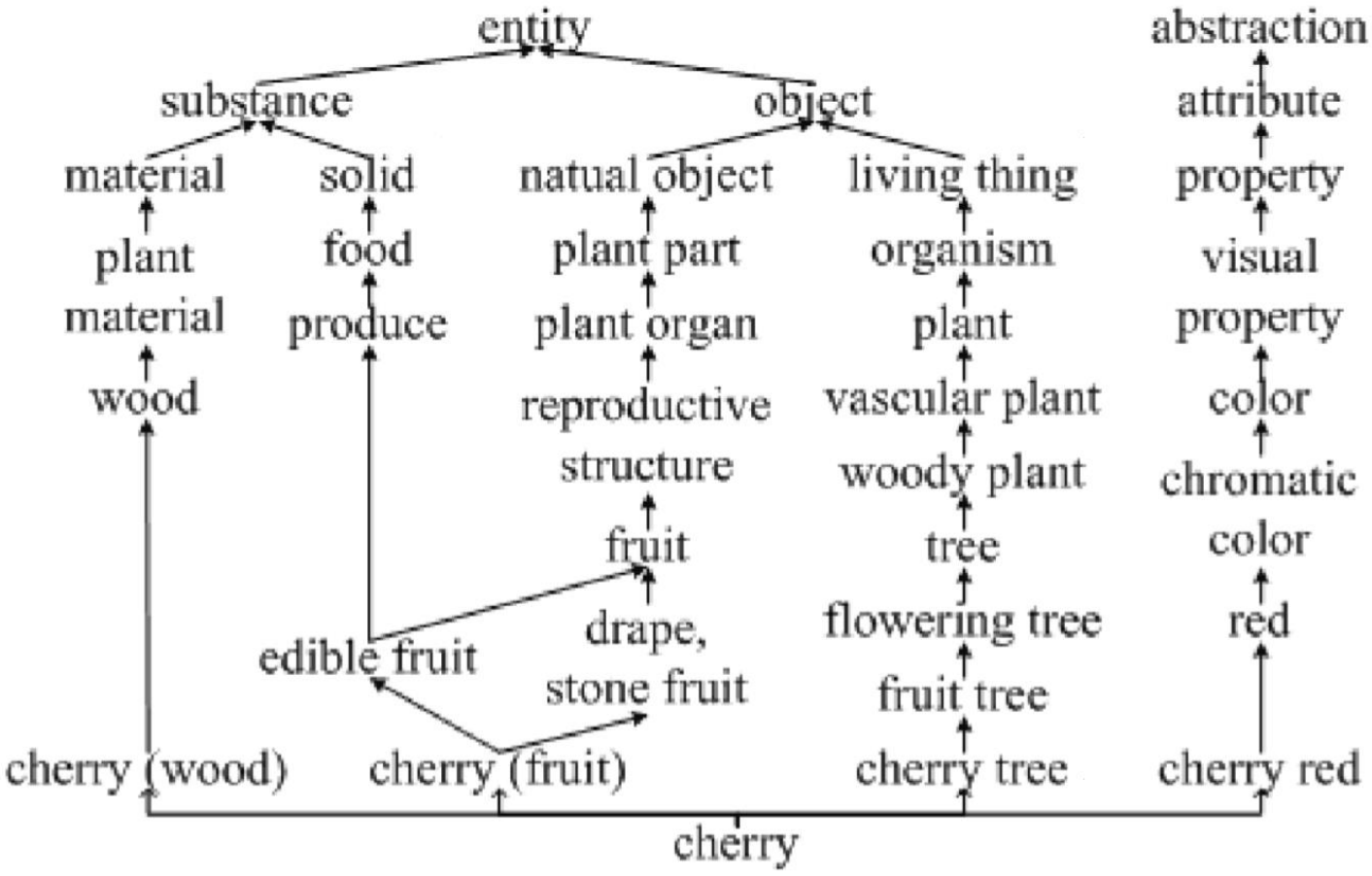
- We have only **raw text**
- Filter out all words but **nouns**
- **Disambiguate** the words
- Map the words to **WordNet concepts**
- We utilise **PageRank**

Word sense disambiguation

- We construct graph **G=(V,E)**
- V are all concepts containing **nouns** in document plus those reachable by hypernym and holonym relations
- E are the **hypernym** and **holonym** relations between V

- *Run PageRank to infer the correct senses*

Word sense disambiguation



Idea: TextRank over concepts?

- TextRank links all **co-occurring words**
- We link all potentially **co-occurring concepts**

- *Add these co-occurrence relations to previous graph and run PageRank*

But there is something wrong…

Top 10 key-concepts from Wikipedia article about data structure

- data, information
- union, labor union, trade union, trades union, brotherhood
- memory, computer memory, storage, computer storage, store, memory board
- phonograph record, phonograph recording, record, disk, disc, platter
- structure, construction
- type
- library
- order
- hashish, hasheesh, haschisch, hash
- phylum

…we do not consider the information content

- Analogy between TF-IDF and our method
- We did only the TF part
- It turns out that the **IDF part** is analogical to **information content***

* P. Resnik, “Using information content to evaluate semantic similarity in a taxonomy,” in Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1, ser. IJCAI’95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 448–453. [Online]. Available: http://dl.acm.org/citation.cfm?id=1625855.1625914

What is the information content?

$$IC(c) = -\log P(c)$$

$$P(c) = \frac{freq(c)}{N}$$

$$freq(c) = \sum_{n \in words(c)} count(n)$$

What is the information content?

$$IC(c) = -\log P(c)$$

$$\begin{aligned} idf(w) &= \log \frac{|D|}{|\forall d \in D : w \in d|} = \\ &= -\log \frac{|\forall d \in D : w \in d|}{|D|} = \\ &= -\log P(w) \end{aligned}$$

http://en.wikipedia.org/wiki/Data_structure	
Not considering information content	Considering information content
– data, information – **union, labor union, trade union, trades union, brotherhood** – memory, computer memory, storage, computer storage, store, memory board – **phonograph record, phonograph recording, record, disk, disc, platter** – structure, construction – type – library – order – **hashish, hasheesh, haschisch, hash** – **phylum**	– data, information – type – array – structure, construction – computer memory unit – record – memory, computer memory, storage, computer storage, store, memory board – class – model, example – queue

Evaluation - text classification

- We used **20 newsgroups** dataset
  - 20 categories of 1000 documents each
- **TF-IDF** as a baseline

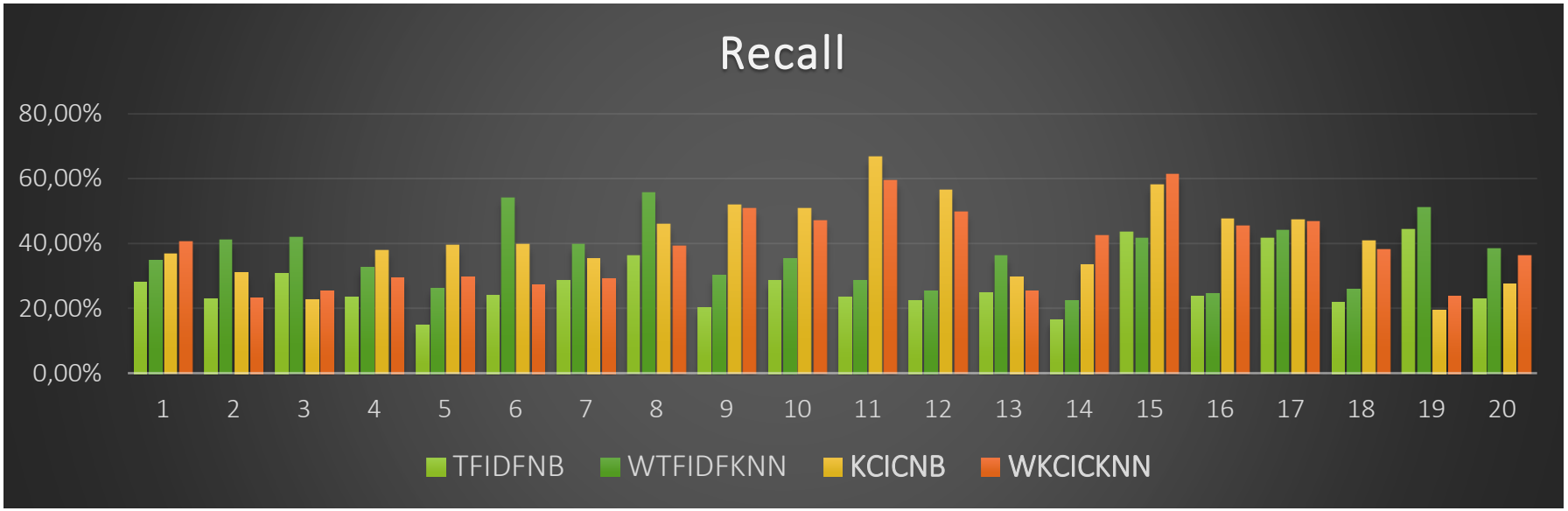
- We represent a document as
  - Top K key-concepts
  - TF-IDF vector
- We use **k-NN** and **Naïve Bayes**

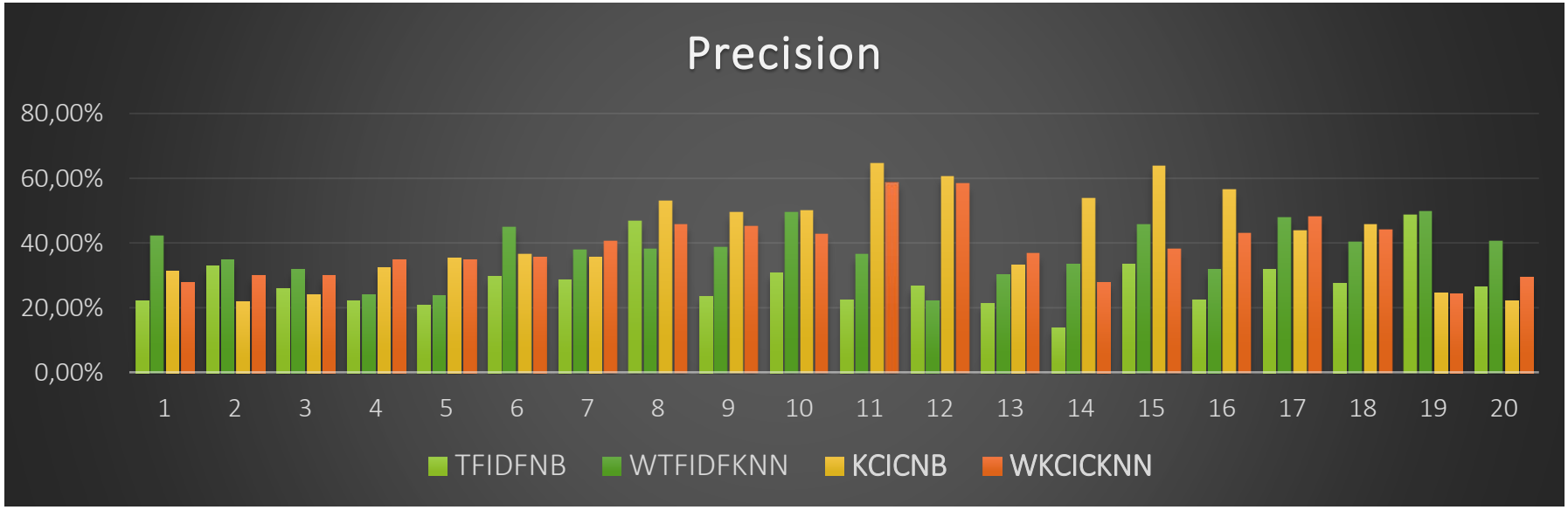
Evaluation - text classification

Method	Accuracy of classification
Top 10 key-concepts with Naïve Bayes	**41.48**
Top 20 weighted key-concepts with k-NN	**38.74**
Weighted TF-IDF vector with k-NN	36.95
TF-IDF vector with Naïve Bayes	27.55

Evaluation - text classification

Number of key-concepts	Accuracy of classification
20	40,77
15	40,73
10	**41,48**
5	40,49
3	38,74
1	29,47


Recall
80,00%
60,00%
40,00%
20,00%
0,00%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
TFIDFNB
WTFIDFKNN
KCICNB
WKCICKNN


Precision
80,00%
60,00%
40,00%
20,00%
0,00%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
TFIDFNB
WTFIDFKNN
KCICNB
WKCICKNN

Conclusion

- A new method of key-concept extraction
- **Key-concepts**
  - Very **efficient**, **concise** representation of document content
  - Easily and **clearly interpretable**
  - Can be used instead of keywords