Sequential patterns for text categorization

Michal Vlasák

Faculty of Informatics Masaryk University

17th December 2013

Problem

- Text categorization

Problem

- Text categorization
- Bag of words

Problem

- Text categorization
- Bag of words
- TF-IDF

Problem

- Text categorization
- Bag of words
- TF-IDF
- Doesn't provide any knowledge

Solution

- Sequential patterns
- More accurate

Solution

- Sequential patterns
- More accurate
- Describe trends
- Provide knowledge

Solution

- Sequential patterns
- More accurate
- Describe trends
- Provide knowledge
- "A customer who bought a TV together with a DVD player, later bought a recorder"

Customers

Based on shopping. Customers buy items in various timespans. Those purchases are represented as sequences.

Purchases

Customer	Date	Items
Peter	12.1.2013	TV (1)
Martin	28.2.2013	Chocolate (5)
Peter	2.3.2013	DVD Player (2), Camera (3)
Peter	12.3.2013	Printer (4)
Peter	26.4.2013	Chocolate (5)

Documents

- Customer → Document
- Item → Word
- Items/transaction → Sentence
- Date → Position of the sentence in document
- Searching for all sequences with $supp(s) \geq minSupp$
- Classification rules are sequences with $confidence(s) \geq minConfidence$
- Ruleset is reduced

Improvements

1. *minSupp* value different for each category
2. Combination with decision trees, NB, etc.
3. Voting for the category

SPaC

- Words represented using TF-IDF
- Removed stop-list words and words with low information gain
- Minimal support automatically computed and changed dynamically
- For each sequence, confidence is determined

Results?