Text analysis by discovering frequent phrases

Helena Ahonen-Myka
Department of Computer Science
University of Helsinki

Many types of phrases can be found in text

- syntactical phrases
  - e.g. noun phrases: ‘a green ball’,
    verb phrases: ‘saw a ball’
- statistical phrases
  - frequent n-grams (consecutive words)
  - frequent word sequences
    - of any length, gaps allowed

A set of documents

1: The Congress subcommittee backed away from mandating specific retaliation against foreign countries for unfair foreign trade practices

2: He urged Congress to reject provisions that would mandate U.S. retaliation against foreign unfair trade practices

3: Washington charged France West Germany the U.K. Spain and the EC Commission with unfair practices on behalf of Airbus

Frequent word sequences

- assume: *S* is a set of documents
  - each document consists of a sequence of words
- a sequence *p* **occurs** in a document *d* if all the words of *p* occur in *d*, in the same order as in *p*
- a sequence *p* is **frequent** in *S* if *p* occurs in at least σ documents of *S*
  - σ is the frequency threshold given
- a **maximal gap** *n* can be given: the occurrences of any two consecutive words of a sequence can have at most *n* words between them

Maximal frequent sequences

- a **maximal frequent sequence**: a sequence of words that
  - appears frequently in the document collection (given a frequency threshold)
  - is not included in another longer frequent sequence

A maximal sequence with subsequences

- dow jones industrial average
  - dow jones
  - dow industrial
  - dow average
  - jones industrial
  - jones average
  - industrial average
  - dow jones industrial
  - dow jones average
  - dow industrial average
  - jones industrial average

Examples of meaningful subsequences

- maximal sequences:
  - rector of the university Ilkka Niiniluoto
  - rector Victor Vasiliev
- subsequences:
  - rector of the university
  - rector Ilkka Niiniluoto
  - Victor Vasiliev

Examples of maximal frequent sequences (Reuters)

- bundesbank president karl otto poehl
- european monetary system ems

- bank england provided money market assistance
- board declared stock split payable april
- boost domestic demand

- expects higher
- expects complete

A long maximal frequent sequence (and one instance)

- federal reserve entered u.s. government securities market arrange repurchase agreements fed dealers federal funds trading fed began temporary supply reserves banking system (22 words)

- The **Federal Reserve entered** the **U.S. Government securities market** to **arrange** 1.5 billion dlrs of customer **repurchase agreements**, a **Fed** spokesman said. **Dealers** said **Federal funds** were **trading** at 6-3/16 pct when the **Fed began** its **temporary** and indirect **supply** of **reserves** to the **banking system.**

Discovery of frequent sequences

- usually a preprocessing phase is needed
  - very common words are removed
  - some punctuation may be removed
  - numbers removed or converted
  - stemming etc. possible
    - countries -> countr

Algorithm 1 (straightforward)

- basic idea: bottom-up approach

  1. Collect all the pairs from the documents, count them, and select the frequent ones
  2. Build sequences of length *k+1* from frequent sequences of length *k*
  3. Select sequences that are frequent
  4. If sequences left: Go to 2

- finally: select maximal sequences

Problems

- frequent sequences in text can be long (more than 20 words)
- processing of all the subsequences of all lengths is not possible
  - bottom-up approach does not work
- restriction of the length?
- higher frequency threshold?

Algorithm 2 (combining bottom-up and greedy approaches)

- frequent pairs are collected
- longer sequences are constructed from shorter sequences (*k*-grams) as in bottom-up approach
- maximal sequences are discovered directly, starting from a *k*-gram that is not a subsequence of any known maximal sequence
- *k*-grams that cannot be used to construct any new maximal sequences are pruned away after each level

Example: original documents

- Rector Ilkka Niiniluoto gave a speech.
- The seminar was opened by Rector of the University of Petrozavodsk, Professor Victor Vasiliev.
- Rector of the University of Helsinki Ilkka Niiniluoto will be present in the meeting.
- Rector of the university Victor Vasiliev signed the agreement.

Example: after removing very common words

- rector ilkka niiniluoto gave speech
- seminar opened rector university pedrozavodsk professor victor vasiliev
- rector university helsinki ilkka niiniluoto present meeting
- rector university victor vasiliev signed agreement

Example: after removing words that occur infrequently

- rector ilkka niiniluoto
- rector university victor vasiliev
- rector university ilkka niiniluoto
- rector university victor vasiliev

- ... assuming frequency threshold 2

Set of frequent pairs:

(rector, ilkka), (rector, niiniluoto), (ilkka, niiniluoto),

(rector, university), (rector, victor), (rector, vasiliev),

(university, victor), (university, vasiliev), (victor, vasiliev)

Expanding a pair:

(rector, ilkka) → (rector, ilkka, niiniluoto) is maximal

Set of frequent pairs:

(rector, ilkka), (rector, niiniluoto), (ilkka, niiniluoto),

(rector, university), (rector, victor), (rector, vasiliev),

(university, victor), (university, vasiliev), (victor, vasiliev)

Expanding a pair:

(rector, university) → (rector, university, vasiliev)

→ (rector, university, victor, vasiliev) is maximal

(rector, ilkka), (rector, niiniluoto), (ilkka, niiniluoto),

(rector, university), (rector, victor), (rector, vasiliev),

(university, victor), (university, vasiliev), (victor, vasiliev)

No more pairs can be expanded.

Maximal frequent sequences after the first pass:

(rector, ilkka, niiniluoto)

(rector, university, victor, vasiliev)

The 3rd level, a set of 3-grams:

(rector, ilkka, niiniluoto), (rector, university, victor),

(rector, university, vasiliev), (rector, victor, vasiliev),

(university, victor, vasiliev)

(rector, ilkka, niiniluoto) is maximal → it can be removed from the set of 3-grams

The 4th level, a set of 4-grams:

(rector, university, victor, vasiliev)

Pruning

- if a *k*-gram is a maximal frequent sequence, it can be pruned
  - a maximal sequence of length *k* can be pruned on the level *k*
  - if there are frequent sequences of length, say 15, this is too late -> we need some ways to prune *k*-grams that cannot be used to build any new frequent sequences

Challenge:
fixed, slightly varying parts

- ”Read more about this topic in the Saturday’s edition of Helsingin Sanomat (www.helsinginsanomat.fi/international)...”
- … in the Monday’s edition…
- … (www.helsinginsanomat.fi/sport)...

Generalized sequences

- the method for discovering the maximal frequent word sequences may be extended to **extract generalized sequences from annotated text**

Generalized sequences

- How the word ’right’ is used? -> concordance of ’right’:

Is that the	**right** time?
...that things weren’t	**right** between us.
Stay	**right** here.
They had the	**right** to strike.

After morphological analysis

- They had the **right** to strike.
  - <they, Pronoun, plural, 3>
  - <had, Verb, past, singular, 3>
  - <the, Determiner>
  - <**right**, Noun, singular,3>
  - <to, Preposition>
  - <strike, Verb, infinitive>

Generalized context of ’right’

the	**right** ’Noun’
be	**right** between ’Pronoun’
’Verb’	**right** here
the	**right** to ’Verb’

Example

- I saw a red ball and a green ball.
- The red ball was small.
- The green ball was big.
- He saw the balls as well.

Document 1:

i	i	nom	pron
saw	see	past	v
a	a	sg	det
red	red	abs	a
ball	ball	nom	n
and	and	nil	cc
a	a	sg	det
green	green	abs	a
ball	ball	nom	n

Representation of sequences

- earlier: a sequence of words
  - i, saw, a, red, ball, and, a, green, ball
- now: a sequence of feature vectors
  - <i,i,nom,pron>, <saw,see,past,v>,...

Discovery of frequent sequences

- occurrences of <saw,see,past,v> is a subset of the occurrences of <see,past,v>
  - saw, see, past, v
  - see, past, v
  - past, v
  - v
- in discovery: if a sequence is not frequent, features are dropped from the feature vectors → generalization

Occurrence of a sequence

- sequence
  - <the, the, nil, det> <abs, a>
    <ball, ball, nom, n> <abs, a>
- occurs in
  - <the, the, nil, det> <red, red, abs, a>
    <ball, ball, nom, n> <was, be, past, v>
    <small, small, abs, a>

Algorithm

- the first step is modified
  - frequent pairs of (subsets of) feature vectors
  - <i, i, nom, pron> <saw, see, past, v>
    - is not frequent
  - <nom, pron> <saw, see, past, v>
    - is frequent, because of ”He saw”, which shares many features

All frequent pairs as input for the 2nd level

<nom, pron> <saw, see, past, v>

<nom, pron> <see, past, v>

<nom, pron> <past, v>

<nom, pron> <v>

<pron> <saw, see, past, v>

<pron> <see, past, v>

<pron> <past, v>

<pron> <v>

Discovery

- as with ”plain” word sequences: feature vectors are treated like words
- resulting maximal frequent word sequences have to be pruned
  - the most specific sequences remain
    - e.g. <nom,pron><saw,see,past,v><det> <ball,nom,n> remains
    - but <pron><saw,see,past,v><det><n> and 22 other subsequences are pruned

<det> <green, green, abs, a> <ball, ball, nom, n>

-- a green ball + the green ball

<det> <red, red, abs, a> <ball, ball, nom, n><abs, a>

-- a red ball green + the red ball small

<the, the, nil, det> <abs, a> <ball, ball, nom, n>

<was, be, past, v> <abs, a>

-- the red ball was small + the green ball was big

<nom, pron> <see, past, v> <det><ball, nom, n>

-- I see a ball + he saw the balls

Applications: how to use the phrases?

- content descriptors (together with single words)
  - for information retrieval (query vs. document similarity)
  - for clustering of documents and labelling of clusters
- text/linguistic analysis
- text summarization
- text compression

Discussion

- the method for discovery of ”plain” word sequences works with rather large document collections (but could be improved)
- the generalized extension would need more efficient data structures and modification of discovery phase