An Automated Research Paper Classification Method for the IPC system with Concept Base for Japanese Subtask at NTCIR-7 Patent Mining Task

Takanori Shimano and Takashi Yukawa
(Nagaoka University of Technology)

Background and Motivation

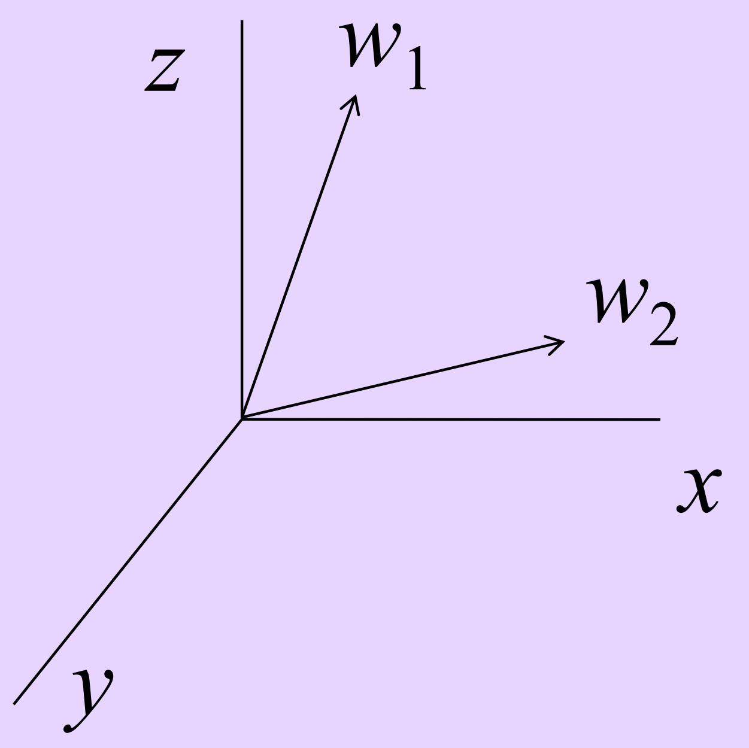
There is a problem of classification in that patent documents differ from research papers respect to document characteristic that is as follows:

- Term
- Document structure

Concept Based Vector Space Model

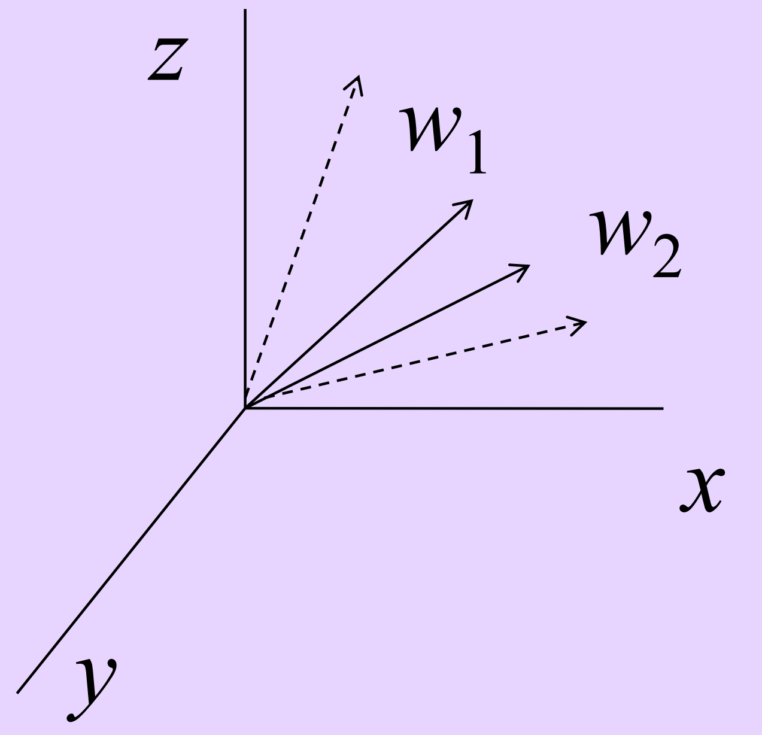
To solve the problem of the term difference, a classification method using the CBVSM is proposed.

In a simple Vector Space Model, vectors of words do not point in the same direction even if they are synonym each other.



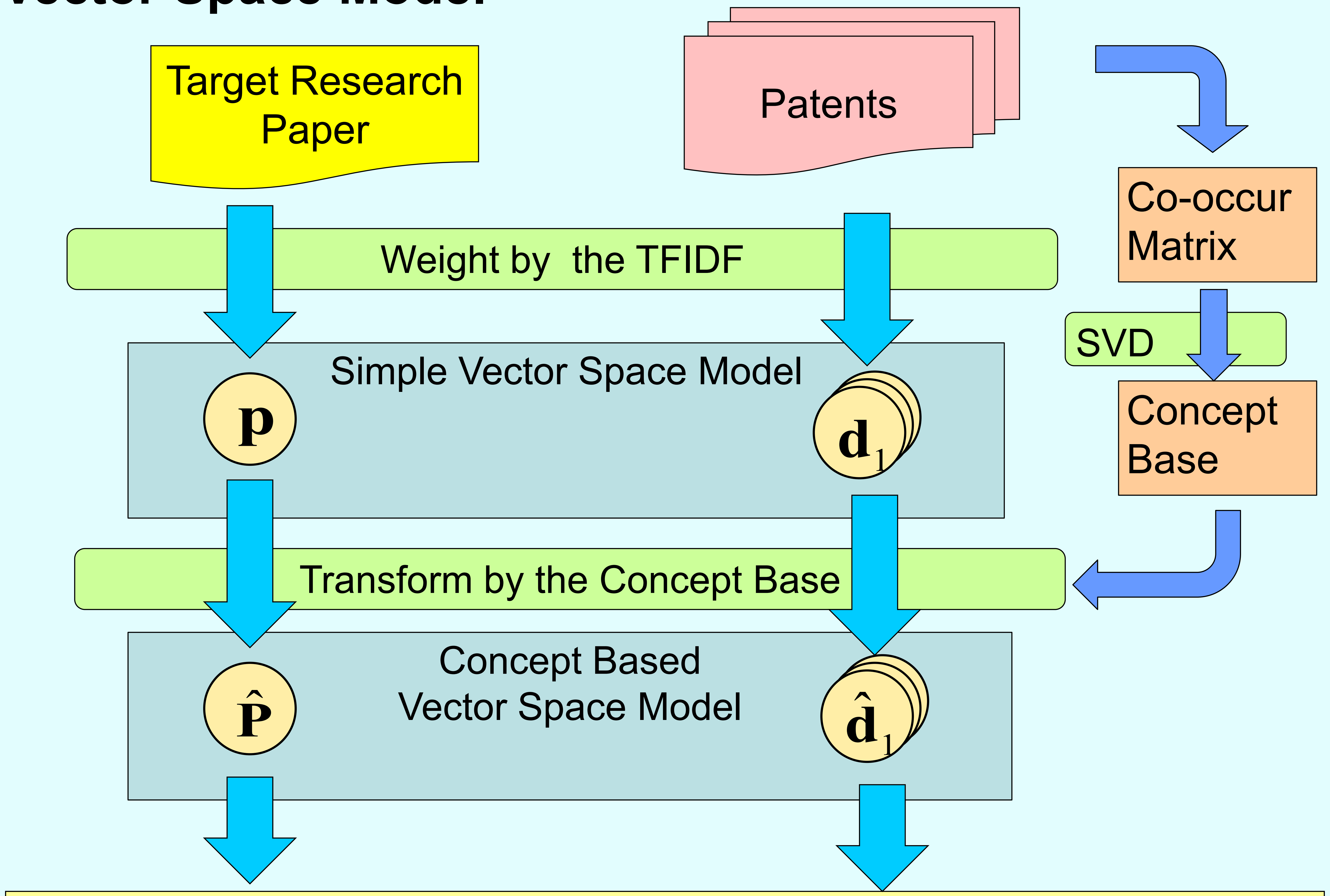
Simple Vector Space Model

In a Concept Based Vector Space Model, vectors of semantically similar words point in the similar direction.



Concept Based Vector Space Model

A Classification Method using the Concept Based Vector Space Model



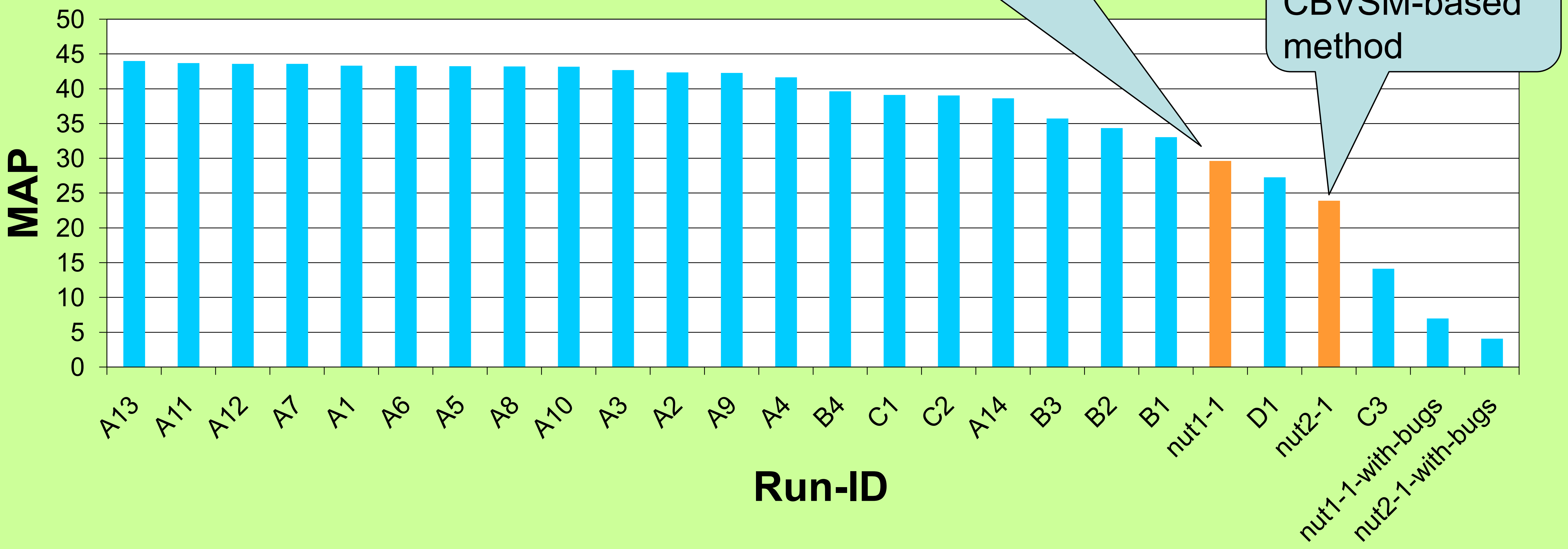
Classification

- Compute the score of the research paper for each class
- Rank the classes based on the score

$$score(\hat{\mathbf{p}}, C) = \sum_{\mathbf{d}\in C} sim(\hat{\mathbf{p}}, \hat{\mathbf{d}}) = \sum_{\mathbf{d}\in C} \frac{\hat{\mathbf{p}} \cdot \hat{\mathbf{d}}}{|\hat{\mathbf{p}}||\hat{\mathbf{d}}|}$$

Evaluation Results

Mean Average Precision



using the simple VSM-based method

using the CBVSM-based method

Comparison of AP values

The AP value of the CBVSM-based method is higher than the simple VSM-based method in **33%** of all topics.

Topic-ID	Simple VSM-based method	CBVSM-based method
300	0.0556	0.1667
301	0.0164	0.0833
302	0.1111	1.0000
303	1.0000	0.0139
…	…	…
MAP	0.2963	0.2388

Discussion

The word that has a too large TF value decreases the MAP value

A Document Vector of Patent #95070321

Word	TF	DF	TFIDF
w_1	7	219657	0.0552
…			
w_6	163	770947	0.7026
…			
w_{20}	1	360676	0.0065

The TF value of the word w_6 is too large, however, the word is not useful for retrieval of patent documents in the topic.

Why the AP value of the CBVSM-based method is higher in 33% of all topics?

In classification of the topic #302, the AP value of the CBVSM-based method is higher than the simple VSM-based method.

Document Vectors on the Simple VSM-based Method

Word	Topic	d_1	d_2
w_7	0.1415	0.0895	0.0127
w_{10}	0.4772	0.5744	0.2741
w_{21}	0.0000	0.1600	0.0000
…			
sim		0.4760	0.5085

Document Vectors on the CBVSM-based Method

Concept	Topic	d_1	d_2
c_{341}	-0.2516	-0.1826	-0.1070
c_{346}	0.0867	0.0867	0.1272
c_{359}	-0.2142	-0.2142	-0.1568
…			
sim		0.8397	0.8367

On the CBVSM-based method, the occurrence of the words w_7 and w_{21} contributes to the concept c_{341} and approximates the vector of the correct patent to the vector of the topic.

The word w_{21} does not increase the score of similarity on the simple VSM-based method.

The patent document d_1 includes the same content as the topic #302.

The patent document d_2 includes a content that differs from the topic #302.

Conclusions

- The CBVSM-based classification method is proposed for the research paper classification.
- The MAP value of the method is lower than the simple VSM-based method.
- However, in 33% of all topics, the AP value of the method is higher than the simple VSM-based method.
- In the future, we intend to investigate two areas of concern:
  - Address the problem of a large TF value by setting a ceiling value.
  - Improve the accuracy of the classification using a combination of the CBVSM and the simple VSM.