Algorithm Merits

Keyword Extraction

Advantages

1. Quality of Keywords
    1. The extracted keywords more uniquely describe the document they came from than standard corpus based methods like TF-IDF.
    2. The extracted keywords are easier to understand for humans
        1. North Korea vs. Korea
        2. Prime Minister Koizumi vs. Koizumi
2. Corpus-based techniques do not work well when we extract keywords from an article that did not come from the corpus.
3. When we use another corpus to calculate IDF statistics with the corpus-based technique we require a very large corpus.
    1. Even when all of Google is used our technique out performs the corpus-based one
    2. Google is impractical to use, because the speed of the algorithm is too slow.

	TOP 10	TOP 3	# 1
Reuters-KANT	98.1%	96.7%	89.2%
Reuters-TF-G	86.1%	85.3%	77.0%
Reuters-Reuters	61.3%	60.6%	54.8%
Reuters-TF	58.5%	57.4%	50.0%
Reuters-Yahoo	55.2%	54.5%	48.8%
Yahoo-KANT	99.8%	99.1%	80.6%
Yahoo-TF-G	95.8%	87.5%	52.8%
Yahoo-Yahoo	67.6%	64.8%	50.2%
Yahoo-Reuters	62.0%	58.3%	42.8%
Yahoo-TF	65.3%	56.0%	36.8%
Sports-KANT	99.2%	99.2%	84.3%
Sports-TF-G	94.4%	92.1%	73.0%
Sports-TF	63.0%	60.6%	45.7%
Sports-Reuters	58.3%	57.5%	42.5%

*KANT is our method, TF-G uses Google as the corpus to calculate IDF

Disadvantages

1. TF-IDF is faster than our approach, because it does not require any morphological analysis.

Category Classification

Advantages

1. Can easily add new categories.
2. Can easily add new training data to categories.
3. No need to keep old training data.
4. Training is simple, quick and only needs positive examples.

Topic Discovery and Classification

Advantages

1. The algorithm can start with 0 topics and 0 training data.
2. Can find new topics.
3. No need to manually tag a corpus for training.