Towards Comment-based Cross-Media Retrieval

Martin Potthast Benno Stein Steffen Becker

Overview

Situation:
Comments can be found all over the Web, and on all kinds and types of objects.

Comments have been found to contain index terms, i.e., commenters describe the commented object to some extent [Potthast 2009 @ SIGIR].

Comparing objects across media types is one of the top challenges in multimedia information retrieval.

Retrieval models for this task, however, are difficult to set up since they are mostly based on low-level features.

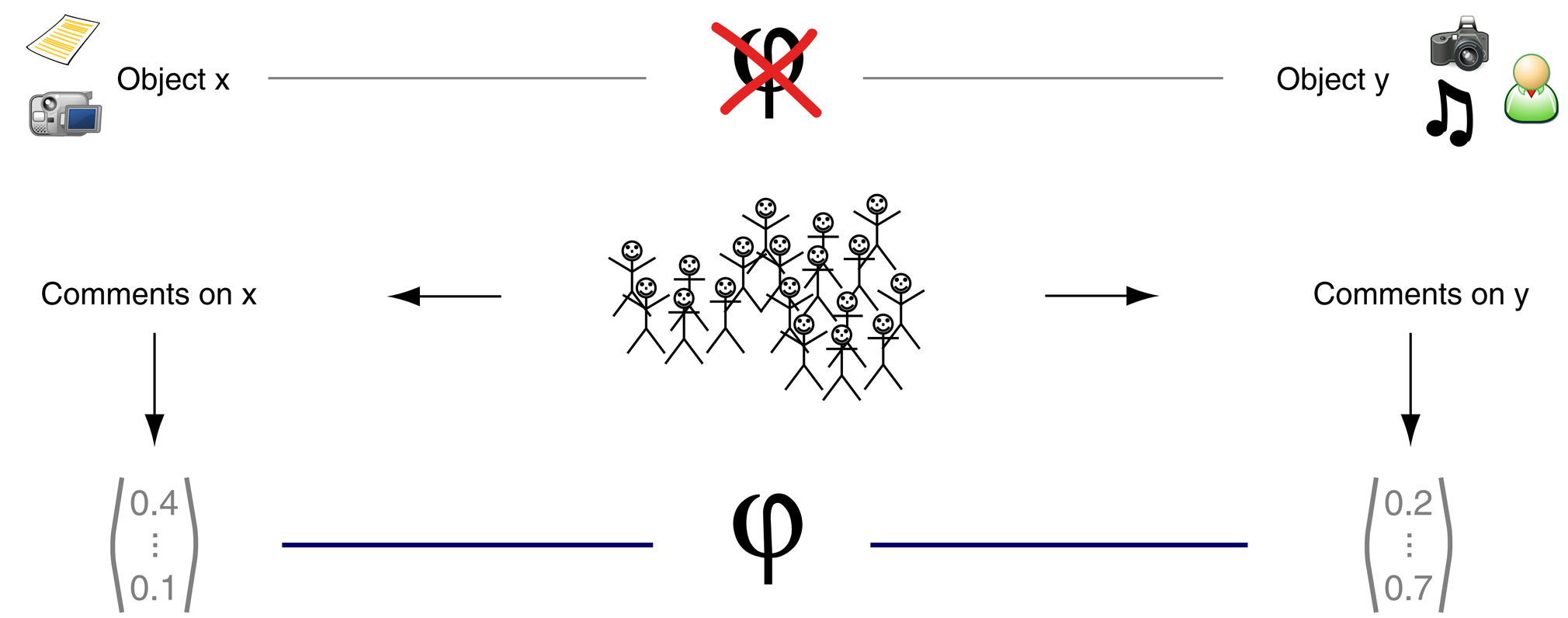
Research question:
Can comments be used to compare the commented objects across media?

Approach:
We use comments to compare news articles from Slashdot with videos from YouTube.

Objects are represented based on their comments using a tfidf-weighted vector space model. As similarity measure the cosine similarity is employed.

Findings:
- 91 out of the top-100 most similar pairs of objects from Slashdot and YouTube are about the same topic.
- Topic matches begin to appear more often at about 0.15 cross-media similarity.
- The simplicity of the retrieval model demonstrates the robustness by which cross-media similarity can be measured based on comments.

Comment-based Cross-Media Retrieval



Evaluation

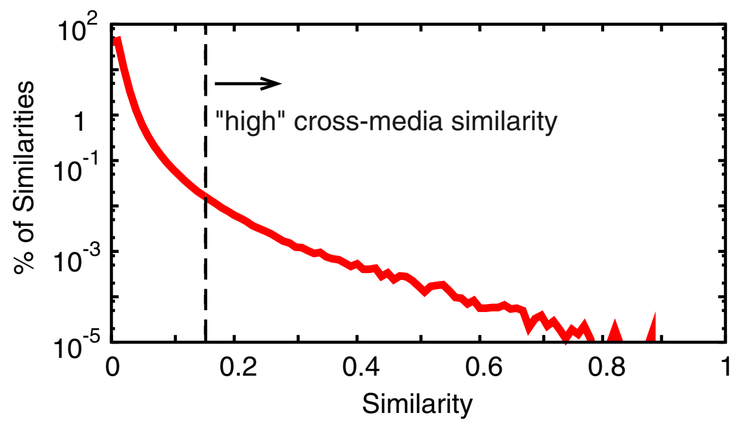
Slashdot compared to YouTube

17 948 articles
1.3 million comments

107.7 million pairs of objects are analyzed using a comment-based vector space model

6 000 videos
4.7 million comments

Distribution of Similarities



Examining the Top-100 Most Similar Pairs

Topic Match	Share	Similarity min	avg.	max	stdev	Avg. # of Comments Slashdot	YouTube	Title Match
equal	36 %	0.71	0.78	0.91	0.06	53	927	72 %
related	55 %	0.71	0.76	0.91	0.04	81	683	62 %
unrelated	9 %	0.72	0.78	0.87	0.05	104	872	--
Σ	100%	0.71	0.77	0.91	0.05	74	790	60 %