A Wikipedia-based Multilingual Retrieval Model
Introducing Cross-language Explicit Semantic Analysis (CL-ESA)

Martin Potthast, Benno Stein, and Maik Anderka
Bauhaus University Weimar
A Wikipedia-based Multilingual Retrieval Model

Introducing Cross-language Explicit Semantic Analysis (CL-ESA)

Explicit Semantic Analysis (ESA) [Gabrilovich:2006].

A document is represented by its similarities to so-called index documents.
In the ESA vector space a similarity is measured using the cosine similarity.

ESA increases the retrieval performance by 20% compared to the VSM.
The ESA model can be applied independently in any language. Is it possible to compare documents from different languages?
A Wikipedia-based Multilingual Retrieval Model

Introducing Cross-language Explicit Semantic Analysis (CL-ESA)

With CL-ESA index documents are chosen from a parallel corpus.

> 100,000 index documents for English and German are available in Wikipedia.
A Wikipedia-based Multilingual Retrieval Model

Introducing Cross-language Explicit Semantic Analysis (CL-ESA)

language-independent vector space
A Wikipedia-based Multilingual Retrieval Model

Introducing Cross-language Explicit Semantic Analysis (CL-ESA)

Wikipedia has more than 250 languages.

Reasonable number of index documents for good performance: 1000 or more.