Related Article Suggestion with Diversity and Relevance Feedback on Turkish Wikipedia Corpus

Devrim Şahin

Motivation and Description

English Wikipedia has **4.5M** articles

Turkish Wikipedia has **230k** articles

Somehow, English Wikipedia is better-organized!

Typos, misclassifications, false information

Difficult to make users provide “Similar Articles”

(Also unreliable)

Motivation and Description

Problem description:

Can we provide an automated list of relevant articles for a given Wikipedia article?

Scope: Turkish Wikipedia

Smaller in size

Data full of noise; a good challenge

Methodology

We want the provided list of items to be diverse:

MMR

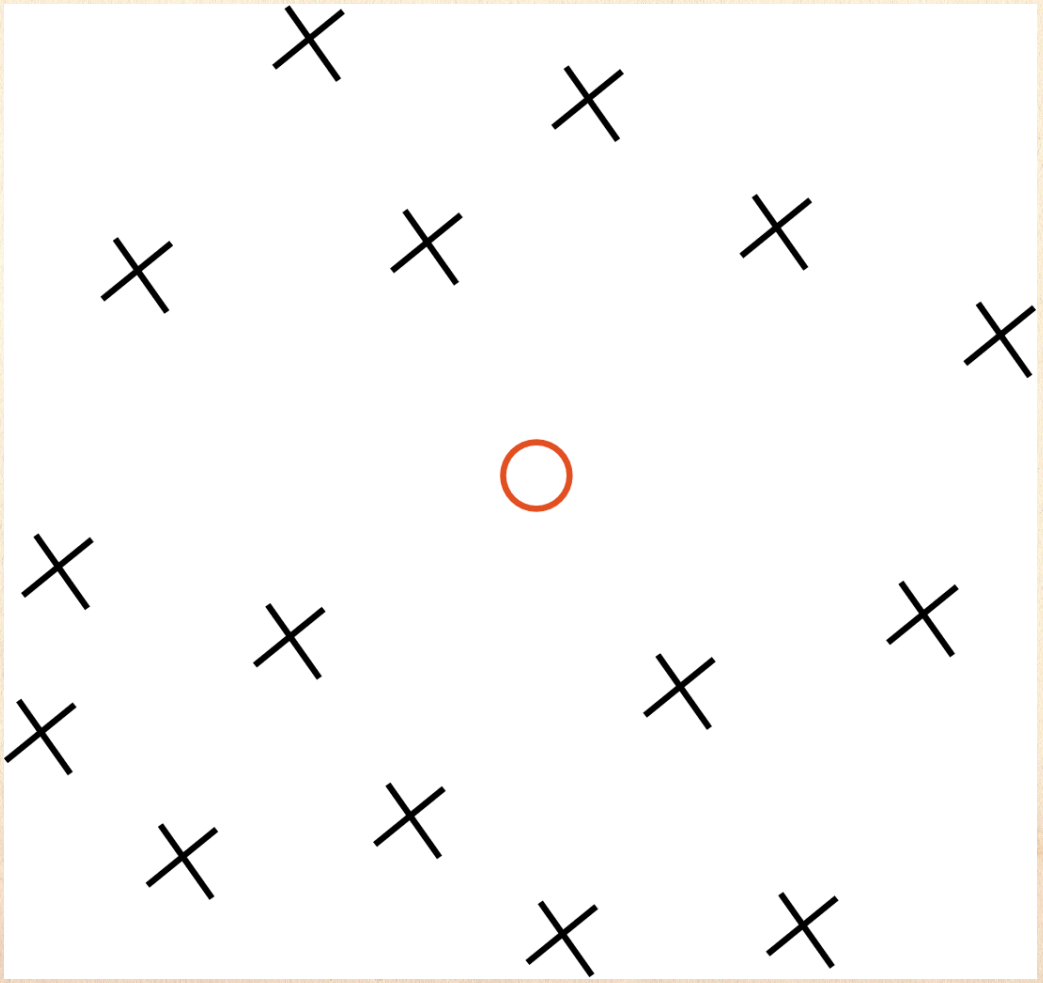
The user might not like the results, and might want to provide feedback:

Relevance Feedback

Instead of Rocchio's algorithm, we will use a novel approach

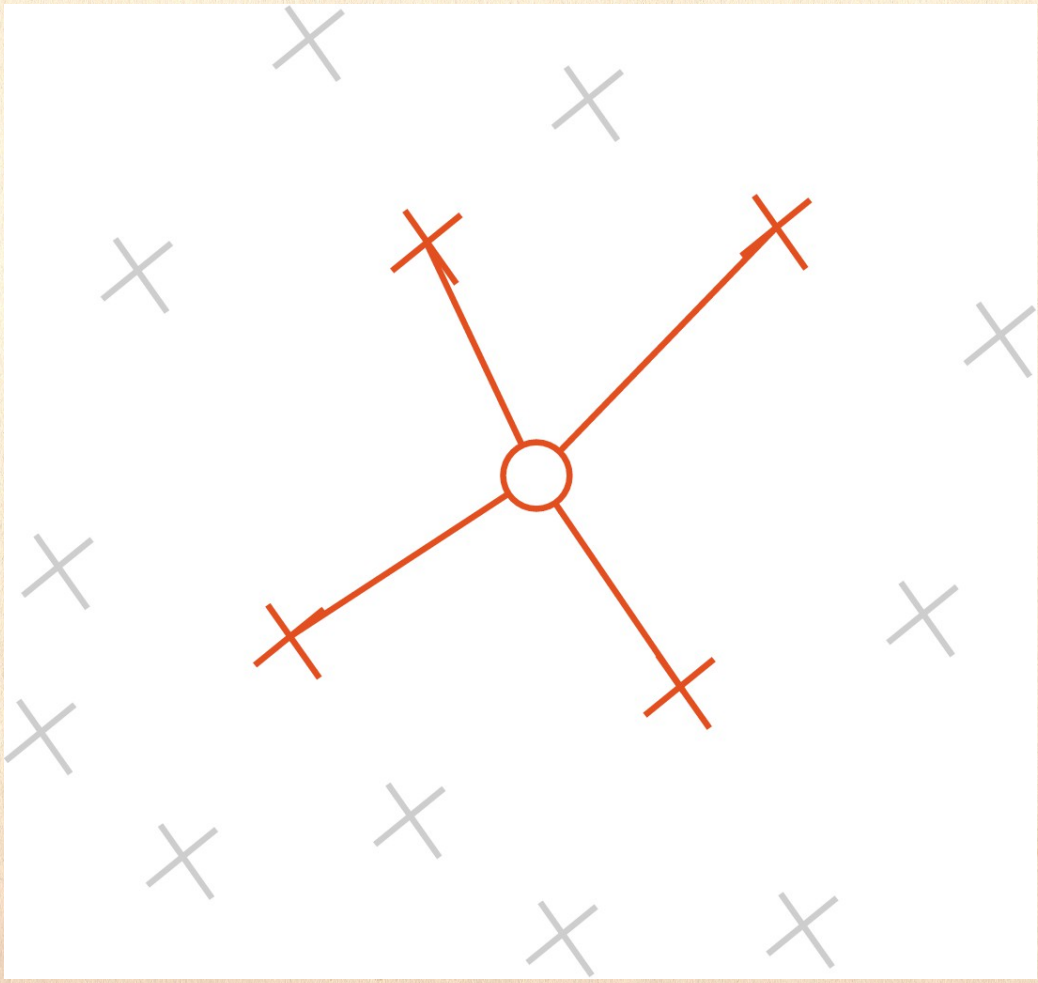
Methodology

Relevance feedback through query expansion



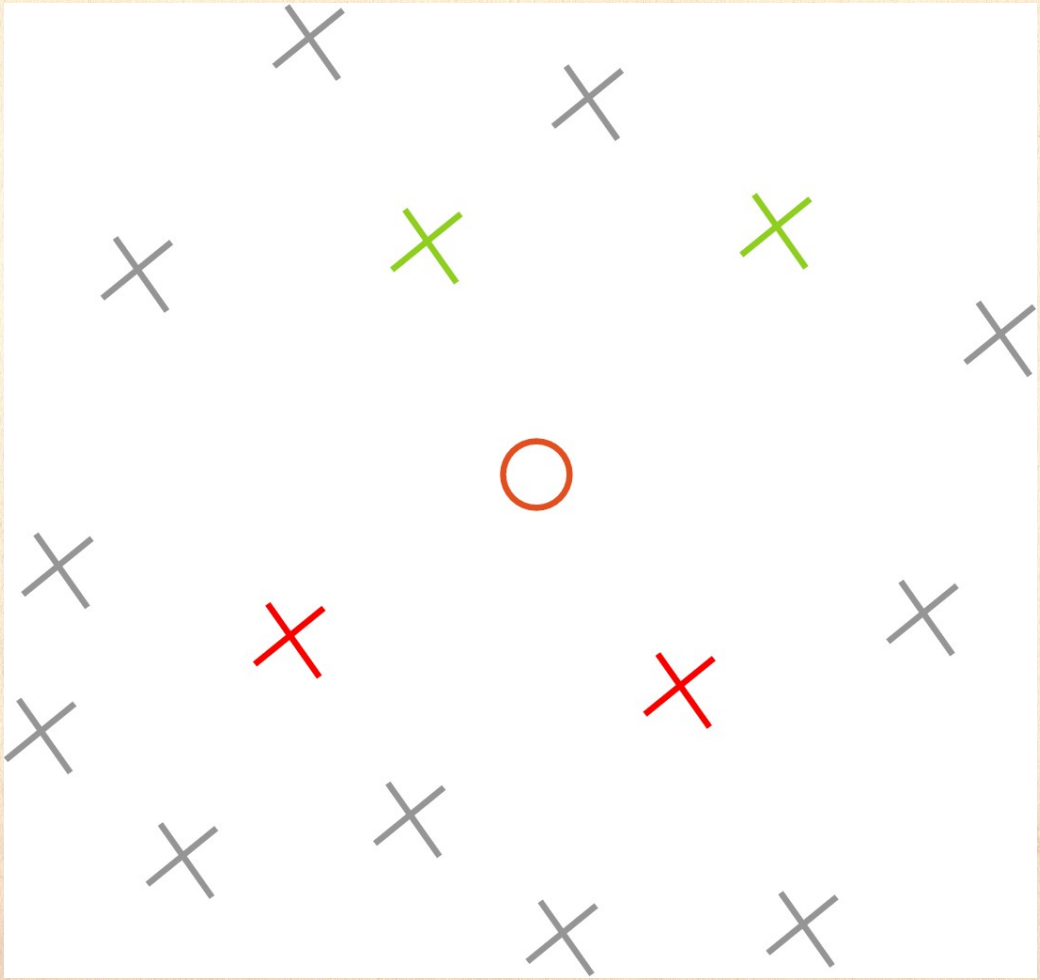
Methodology

Relevance feedback through query expansion



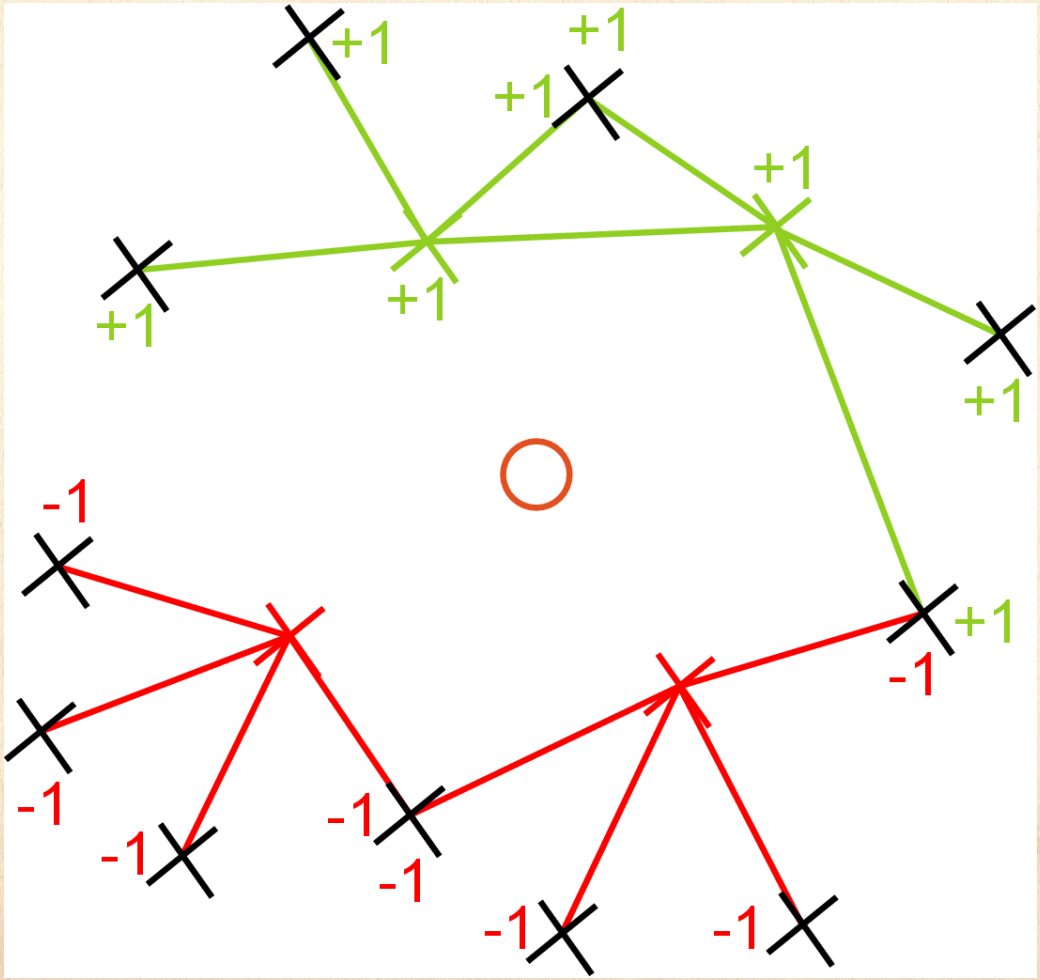
Methodology

Relevance feedback through query expansion



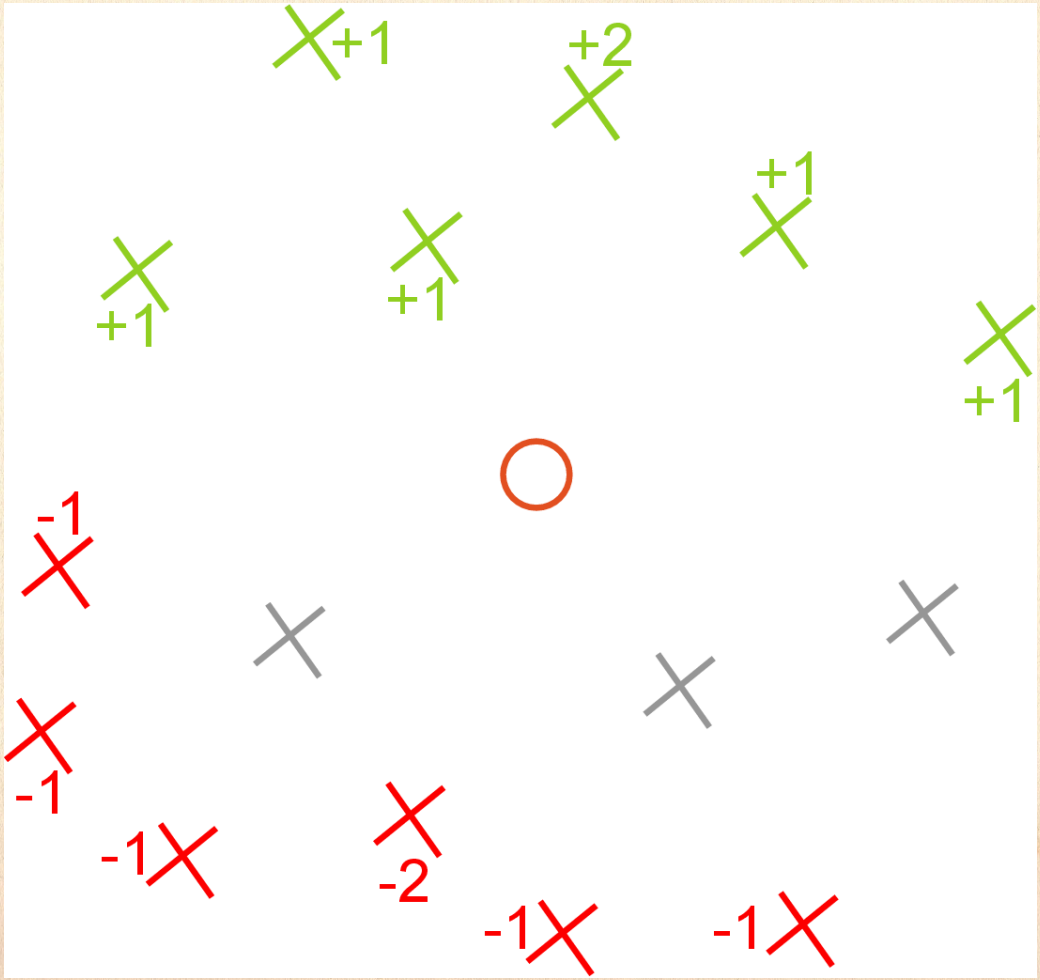
Methodology

Relevance feedback through query expansion



Methodology

Relevance feedback through query expansion



Challenges

The first inspection reveals that Turkish Wikipedia is full of typos

Copy-paste makes it worse

Stemming and term clustering

Size of the corpus

Not as significant for the Turkish Wikipedia

Evaluation

Evaluation

Lack of ground truth for result set relevance

Accuracy and entropy measures can be used for representative vectors

A pooling approach similar to that of the MMR paper can be employed

Thank you for listening

Questions & Comments

References

- Jaime Carbonell and Jade Goldstein. “The use of MMR, diversity-based reranking for reordering documents and producing summaries.” 1998
- Michael E Houle et al. “Can shared-neighbor distances defeat the curse of dimensionality?” In: Scientific and Statistical Database Management. Springer. 2010, pp. 482–500.
- TSCorpus <http://www.tscorpus.com>
- Bahaeddin Eravci and Hakan Ferhatosmanoglu. “Diversity based Relevance Feedback for Time Series Search”. In: Proceedings of the VLDB Endowment 7.2 (2013).