Supporting the Negation Operator in the Hermes Graphical Query Language for Document Ranking

Arnout Verheij, Allard Kleijn, Flavius Frasincar,
Damir Vandic, Frederik Hogenboom

ERASMUS UNIVERSITEIT ROTTERDAM

Introduction

- Hermes serves personalized news to users.
- Users of the Hermes framework use the Hermes Graphical Query language (HGQL).
- HGQL allows users to create fairly complex queries in an intuitive way with little understanding of query languages.
- As HGQL supports negation, we propose a ranking algorithm that is able to deal with negations.

HGQL Ranking

- Like other ranking models, requires a document and query term weight computation method.
- Based on the p-norm Extended Boolean model:

$$\mathrm{sim}\left(\mathbf{d},\mathbf{q}\ \mathrm{OR}_{(p)}\right)=\left(\frac{\sum_{k=1}^{m}(q_k)^p(d_k)^p}{\sum_{k=1}^{m}(q_k)^p}\right)^{1/p}$$

$$\mathrm{sim}\left(\mathbf{d},\mathbf{q}\ \mathrm{AND}_{(p)}\right)=1-\left(\frac{\sum_{k=1}^{m}(q_k)^p(1-d_k)^p}{\sum_{k=1}^{m}(q_k)^p}\right)^{1/p}$$

- This assumes that: if q=[1,1], then:
  - d=[1,1] is ***most*** relevant if q is conjunctive;
  - d=[0,0] is ***least*** relevant if q is disjunctive.
- Change term weight assignment to support the negation operator:
  - In document vectors: -1 instead of 0 if term does not occur;
  - In query vectors: 0 if a term is not part of the query, multiplied with -1 if the term is negated.
- The ranking formulas change as following:

$$\mathrm{sim}\left(\mathbf{d},\mathbf{q}\ \mathrm{OR}_{(p)}\right)=\left(\frac{\sum_{k=1}^{m}(q_k)^p(d_k+q_k)^p}{\sum_{k=1}^{m}(2\times q_k)^p}\right)^{1/p}$$

$$\mathrm{sim}\left(\mathbf{d},\mathbf{q}\ \mathrm{AND}_{(p)}\right)=1-\left(\frac{\sum_{k=1}^{m}(q_k)^p(q_k-d_k)^p}{\sum_{k=1}^{m}(2\times q_k)^p}\right)^{1/p}$$

- This assumes that: if q=[1,1], then:
  - d=[1,1] is ***most*** relevant if q is conjunctive;
  - d=[-1,-1] is ***least*** relevant if q is disjunctive.
- The normalization factor has also been updated.

Evaluation

- We considered the following term weight computation methods:
  - **Rank (eB)**: the extended Boolean model (uses a simple binary weight);
  - **Rank (tfc.tfc)**: the traditional TF.IDF model;
  - **Rank (lxc.ltc)**: the TF.IDF model with logarithmic weights;
  - **Rank (Lnu.ltu)**: the TF.IDF model with document length normalization.
- 5 test users evaluated 927 retrieved news items.
- 'Mean Precision @ 10' (MP@10) and Mean Average Precision (MAP) are used as measures.
- The adapted p-norm Extended Boolean model performs best with a MP@10 of 0.85 and a MAP of 0.87.
- The lxc.ltc algorithm is second best, with an MP@10 of 0.73 and a MAP of 0.694.
- The Lnu.ltu algorithm and the tfc.tfc algorithm achieved an MP@10 of 0.48 and 0.47, respectively, and a MAP of 0.572 and 0.467, respectively.

Conclusion

- We have introduced the negation operator for the p-norm Extended Boolean model in HGQL.
- Our approach uses a negative values in term weights to support negation.
- We have compared different term weighting procedures in our evaluation.
- We conclude that our proposed ranking model works best using binary weights.

Contact

Arnout Verheij:	308057av@student.eur.nl
Allard Klein:	303118ak@student.eur.nl
Flavius Frasincar:	frasincar@ese.eur.nl
Damir Vandic:	vandic@ese.eur.nl
Frederik Hogenboom:	fhogenboom@ese.eur.nl

12th Dutch-Belgian Information Retrieval Workshop