Bing-CSF-IDF+: A Semantics-Driven Recommender System for News

Lies Hooft van Huijsduijnen, Thom Hoogmoed, Geertje Keulers, Edmar Langendoen, Sanne Langendoen, Tim Vos, Frederik Hogenboom, Flavius Frasincar, and Tarmo Robal
Agenda

• Introduction
• Background
  • From TF-IDF to SF-IDF+, CD-IDF+, ..., and Bing-CSF-IDF+
• Bing-CSF-IDF+ recommender
• Evaluation
• Conclusions
Introduction

• Information overload
  • Need for automated and accurate approach in Web to distinguish relevant and non-relevant

• **Recommender systems** (RS) help users to plough through a massive and increasing amount of information
  • Content-based, Collaborative filtering, Hybrid

• Automatically find relevant content based on user preferences, profiles, behaviour

• Meaning of text: semantic lexicon (WordNet) & word sense disambiguation
Introduction: TF-IDF concept

• Common measure: Term Frequency – Inverse Document Frequency (TF-IDF)  
  (Salton and Buckley, 1988)

• Pre-processed documents (stop words removal and stemming)
• For each term, it considers:
  • The importance in a single document
  • The inverse of the general importance within a set of documents

• Users’ interests translated into vectors of TF-IDF weights
• Weights computed for every term within a document

• Similarity between unread news items and user’s interest based on cosine similarity
• Recommendation if similarity > cut-off
Introduction

**TF-IDF**
- **Terms**
  - build on
  - SF-IDF
    - Synsets
    - (27 weights, e.g., antonymy, synonymy)
  - extend
  - SF-IDF+
    - Synsets + Relationships
    - (27 weights: superclass, subclass, domain rel.)
  - extend
  - Bing-SF-IDF+
    - Named entities

**CF-IDF**
- **Concepts**
  - build on
  - extend
  - CF-IDF+
    - Concepts + Relationships
    - (3 weights: superclass, subclass, domain rel.)
  - extend
  - Bing-CSF-IDF+
    - Named entities
    - Concepts + Relationships
    - Synsets + Relationships

Bing

24th European Conference on Advances in Databases and Information Systems (ADBIS 2020)
Bing-CSF-IDF+

- Order of importance:
  1. Concepts and relationships (CF-IDF+)
  2. Named entities with Bing (Bing)
  3. Synsets and relationships (SF-IDF+)

\[
\text{sim}_{\text{Bing-CSF-IDF}^+}(d_u, d_r) = \alpha \times \text{sim}_{\text{Bing}}(d_u, d_r) \\
+ \beta \times \text{sim}_{\text{CF-IDF}^+}(d_u, d_r) \\
+ (1 - \alpha - \beta) \times \text{sim}_{\text{SF-IDF}^+}(d_u, d_r)
\]

- Relies on the Hermes framework (RSS, NLP)
  - Indexing, querying, recommending news items
Evaluation (1)

- Weight optimization for cut-off values
  - 32 weights (27 SF-IDF+ relations, 3 CD-IDF+ relations, and $\alpha$, $\beta$)
  - Genetic algorithm
  - Lisa system computer cluster from SURFsara

- Experiment: same data-splits used for all recommenders
  - Recommenders evaluated: Bing-CSF-IDF+, Bing-SF-IDF+, CF-IDF+
Evaluation (2)

- CF-IDF+ performs well for low cut-off values

- Bing-CSF-IDF+ and Bing SF-IDF+ perform notably better than CF-IDF+

- Bing-CSF-IDF+ has good performance & outperforms others for cut-off values 0.05-0.4
Evaluation (3)

- Bing-SF-IDF+ has low performance for low cut-off values
- Bing-CSF-IDF+ outperforms the other recommenders
Conclusions

• New recommender system combining best features of earlier work: Bing-CSF-IDF+

• Bing-CSF-IDF+ outperforms CF-IDF+ and Bing-SF-IDF+

• Future work:
  • Optimization of weights, e.g., Ant Colony Optimization
  • Include a larger collection of relationships
  • Evaluate on a larger set of news items
Thank you!