Erasmus
School of
Economics

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

Erasmus University Rotterdam
Tallinn University of Technology

TAL TECH

Erasmus

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

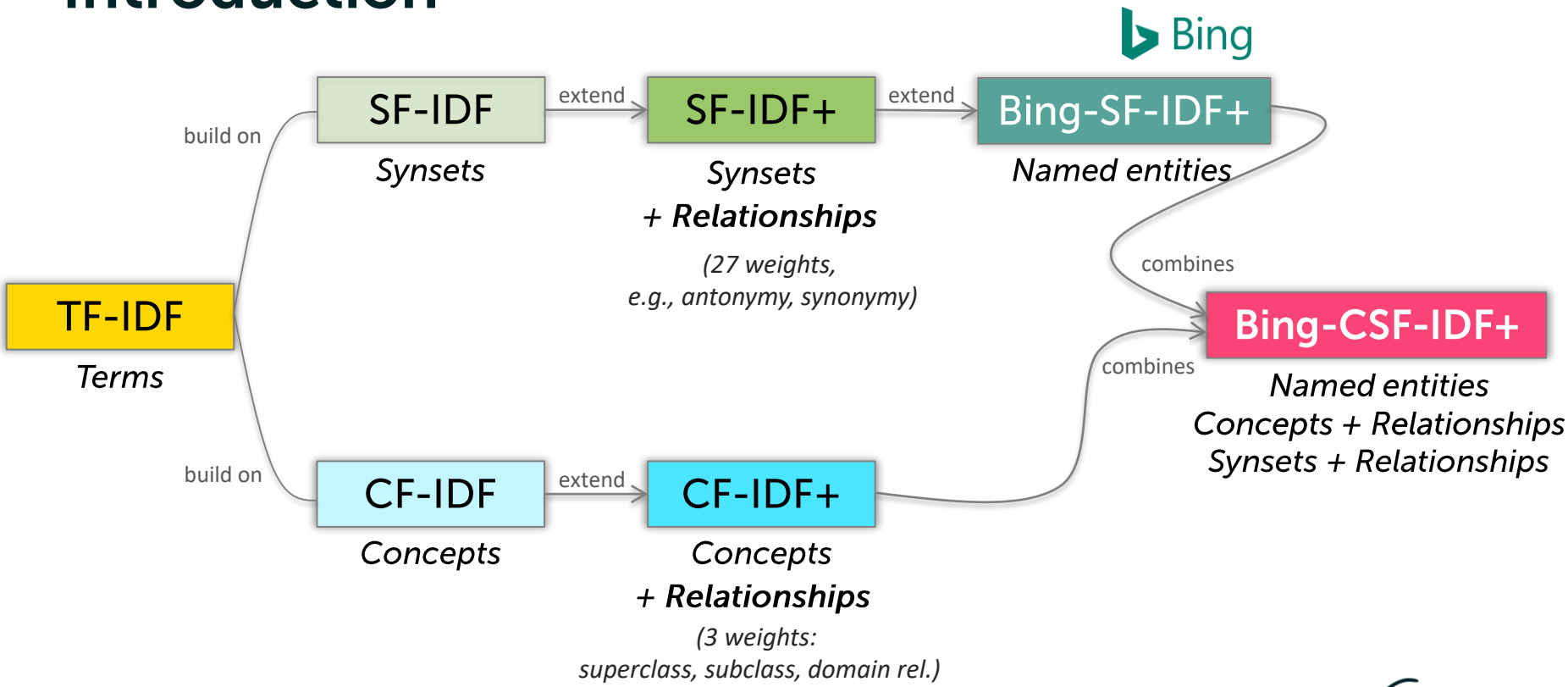
→ Important terms:
red, purple, blue

→ Irrelevant terms:
yellow, green, pink

- 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



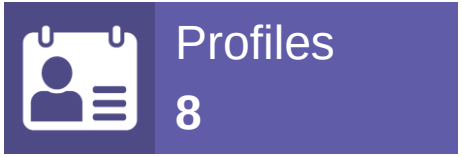
Bing-CSF-IDF+

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

$$\begin{aligned}\mathrm{sim}_{\text{Bing-CSF-IDF+}}(d_u, d_r) = \alpha \times \mathrm{sim}_{\text{Bing}}(d_u, d_r) \\ + \beta \times \mathrm{sim}_{\text{CF-IDF+}}(d_u, d_r) \\ + (1 - \alpha - \beta) \times \mathrm{sim}_{\text{SF-IDF+}}(d_u, d_r)\end{aligned}$$

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

Evaluation (1)

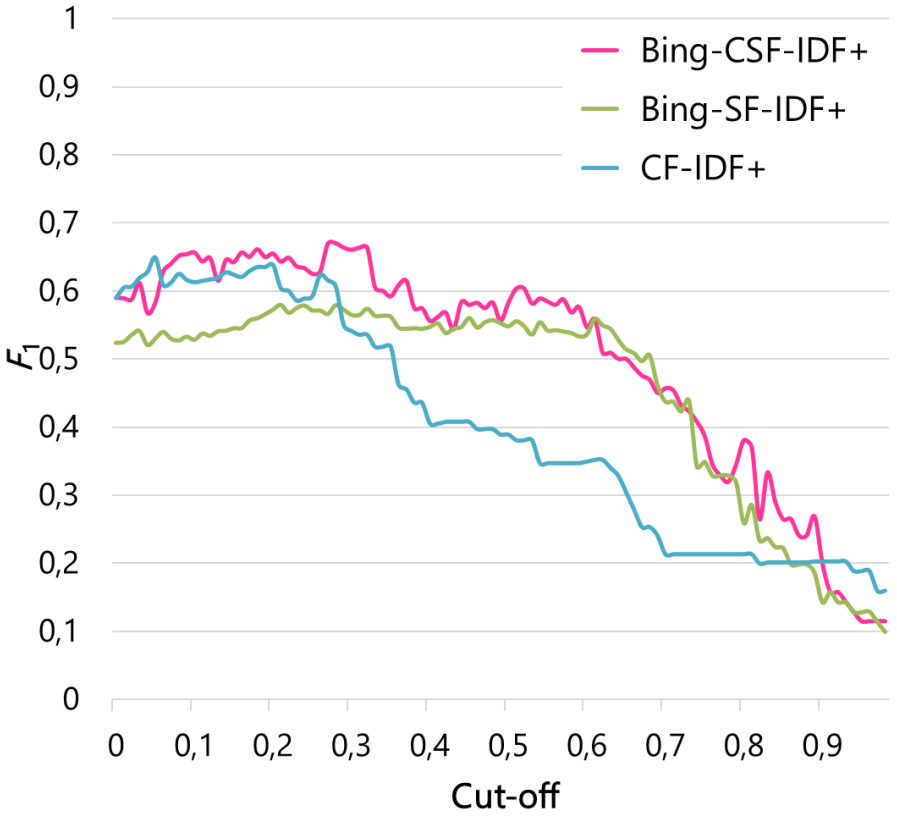


VALIDATION 30%	TRAINING 30%	TEST 40%

- Weight optimization for cut-off values
  - 32 weights (27 SF-IDF+ relations, 3 CD-IDF+ relations, and α, β)
  - 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+

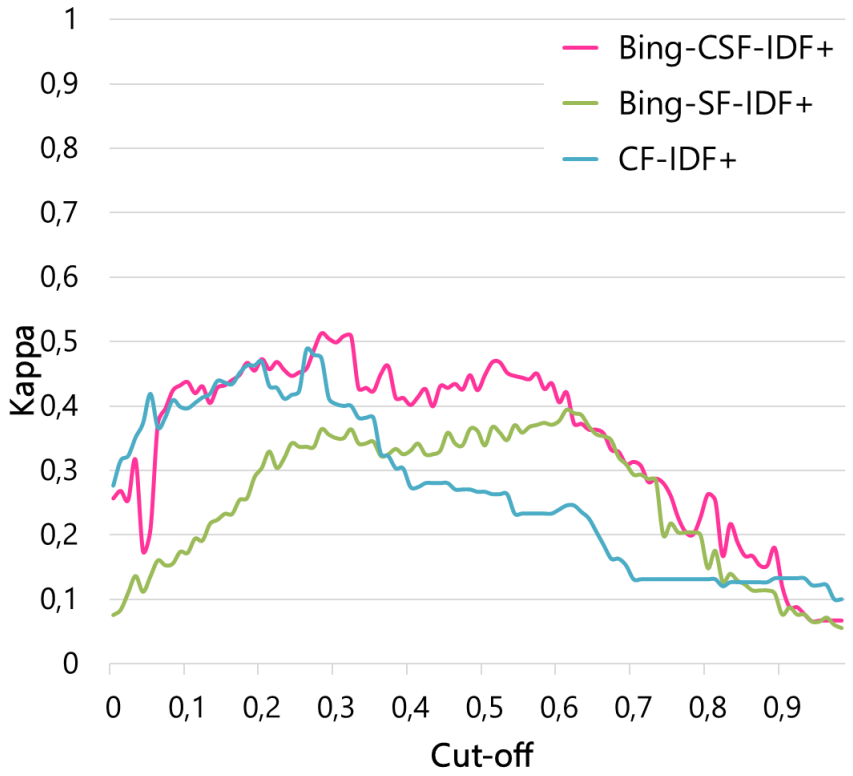
Evaluation (2)

- 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!

Erasmus University Rotterdam
Tallinn University of Technology