Extending FolkRank with Content Data

Nikolas Landia
University of Warwick
Coventry CV4 7AL
UK
N.Landia@warwick.ac.uk

Sarabjot Singh Anand
University of Warwick
Coventry CV4 7AL
UK
S.S.Anand@warwick.ac.uk

Andreas Hotho
University of Würzburg
Am Hubland
Würzburg
Germany
hotho@informatik.uni-wuerzburg.de

Robert Jäschke
University of Kassel
Wilhelmshöher Allee 73
34121 Kassel
Germany
jaeschke@cs.uni-kassel.de

Stephan Doerfel
University of Kassel
Wilhelmshöher Allee 73
34121 Kassel
Germany
doerfel@cs.uni-kassel.de

Folke Mitzlaff
University of Kassel
Wilhelmshöher Allee 73
34121 Kassel
Germany
mitzlaff@cs.uni-kassel.de

Summary

- Extension of FolkRank with content data
- Simpler content-based recommender: WordTags
- Analysis of edge weighting scheme of FolkRank

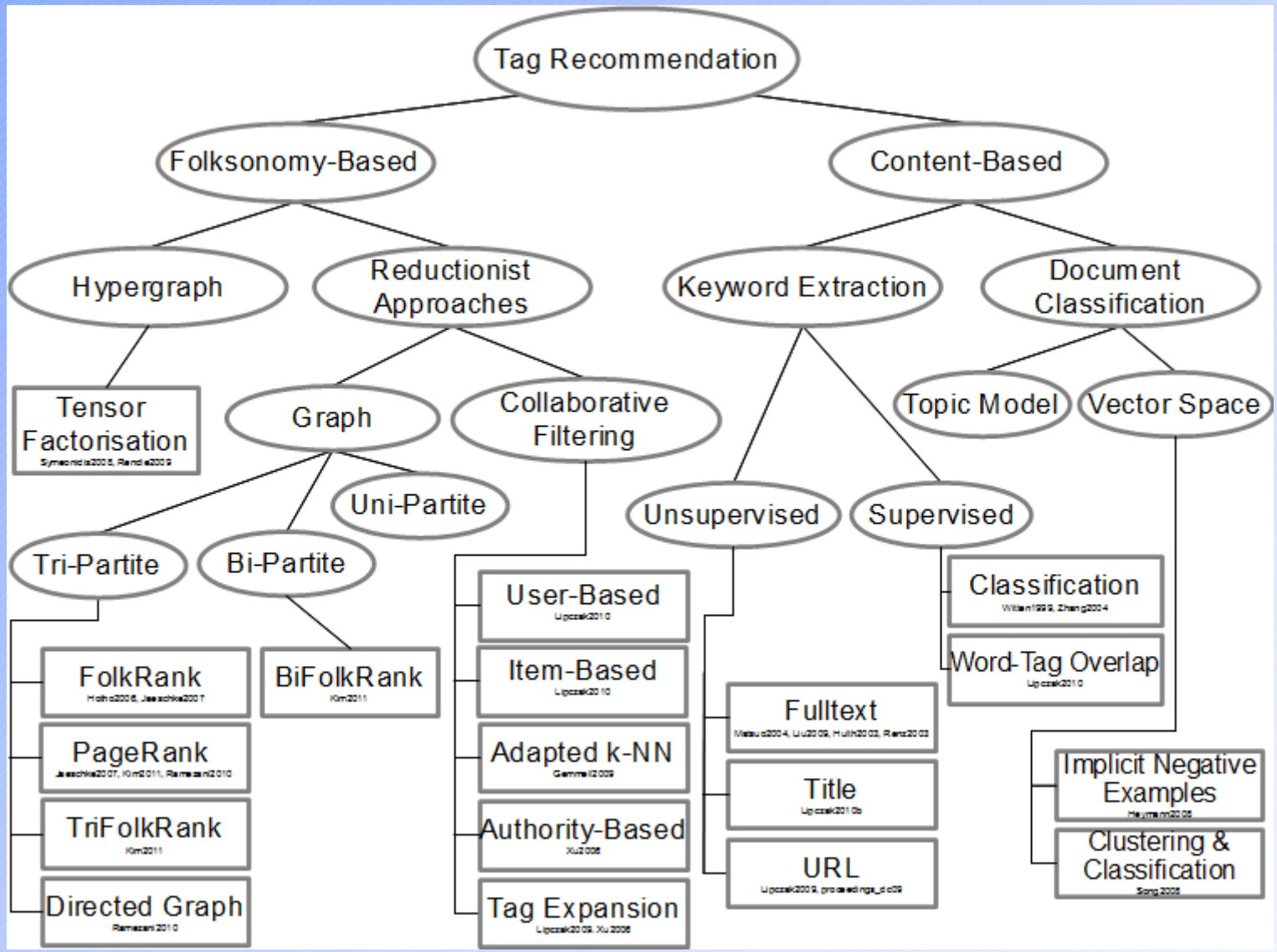
Introduction

- Tagging is a popular document organisation methodology
- Applications include social bookmarking websites such as BibSonomy, CiteULike and Delicious
- Users have the liberty of assigning any string of characters as a tag to a document

Introduction

- A Folksonomy is a collection of tag assignments of the form (user, document, tag) with timestamps
- A “post” is the set of all tag assignments related to a unique (user, document) pair
- Tag Recommendation is the task of suggesting a set of tags to the user for a document that he is in the process of tagging

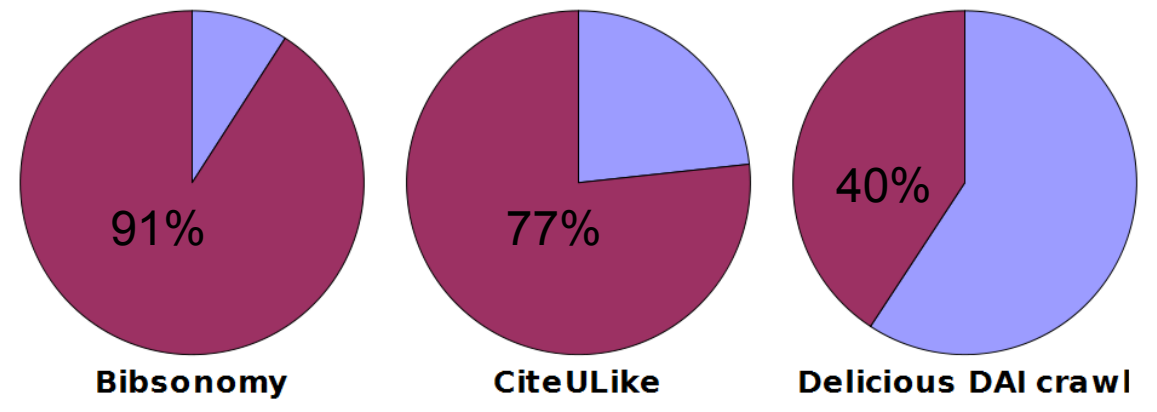
Overview of existing tag recommendation approaches



Why is content important?

- The new item problem with regard to documents is very prominent as most documents are only tagged by one user

Percentage of posts with new documents in social bookmarking datasets



Document Model

- Bag-of-words representation
- Each document is a vector of Tf-Idf scores
- Content sources
  - Title
  - Meta-data: *title, url, author, description, abstract ...*

FolkRank Overview

- Folksonomy-based tag recommender
- Iterative weight spreading algorithm similar to PageRank

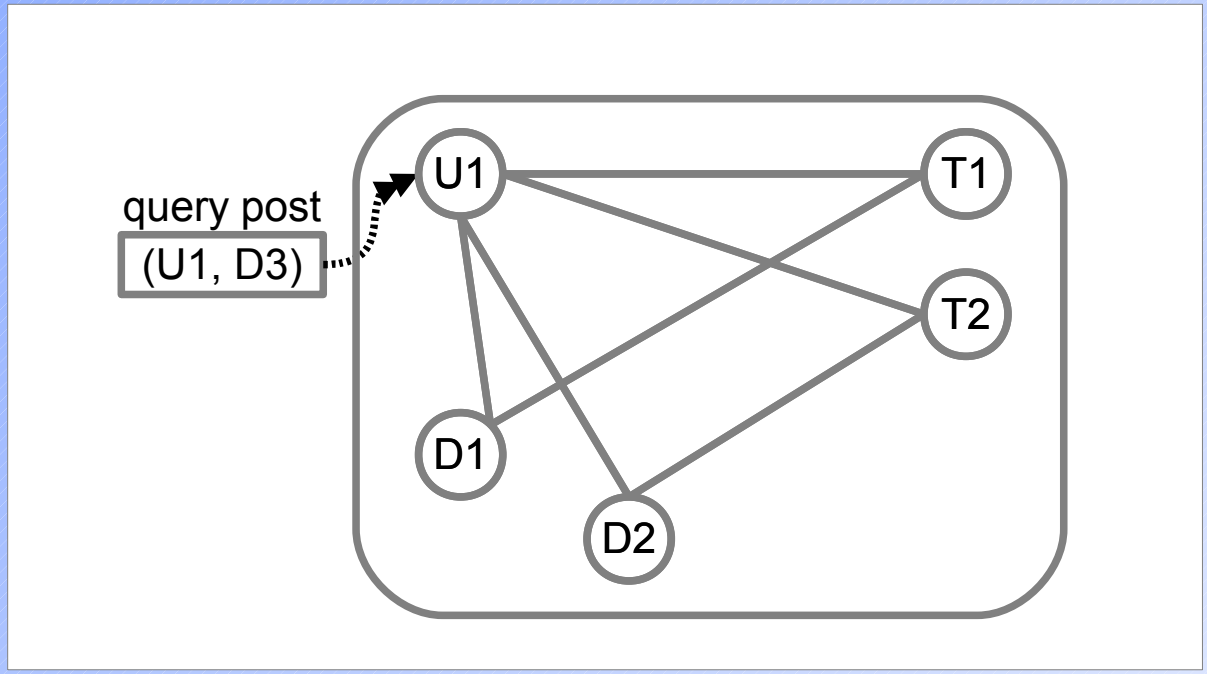
Learning model

- Construct graph which models user, document and tag relationships

Recommendation

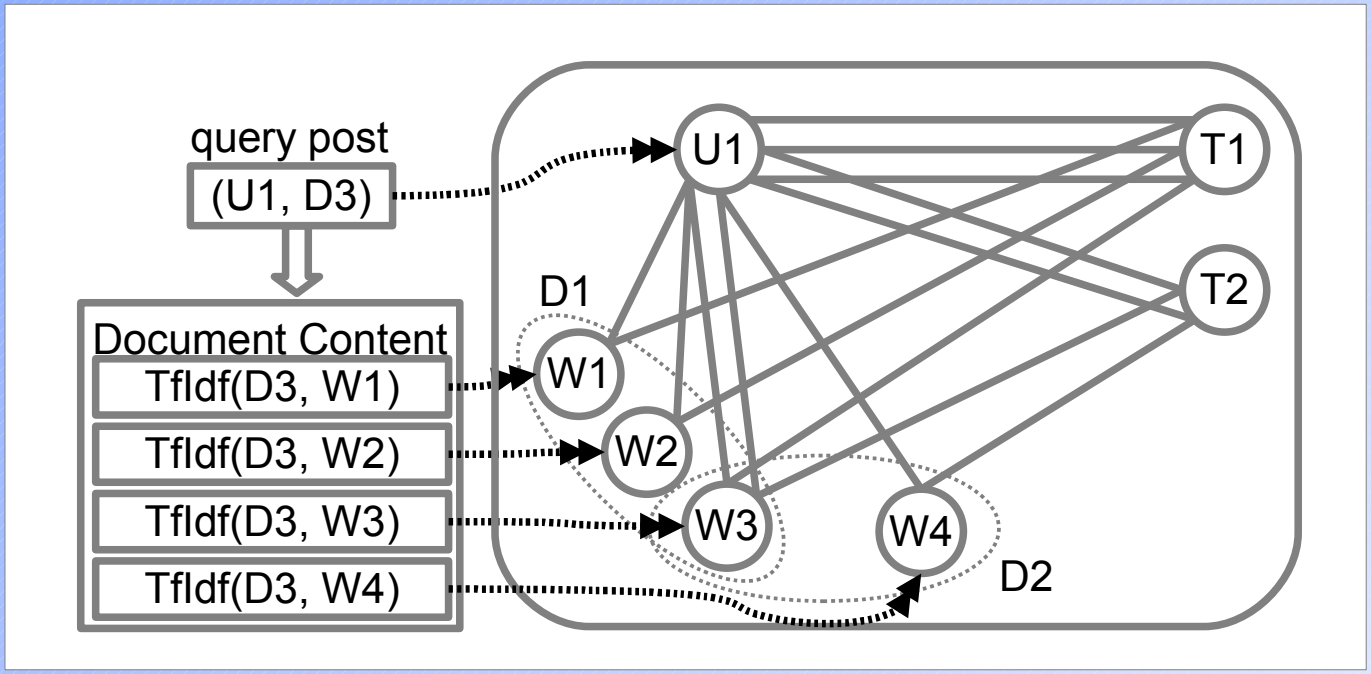
1. Give high preference weight to query user and document
2. Perform weight spreading iterations
3. Stop when node weights stabilise
4. Recommend tags ranked by their weight in graph

FolkRank



- User, document and tag nodes
- Edge weights based on co-occurrence data
- Preference vector consists of query user and query document (if it exists in graph)

ContentFolkRank



- User, word and tag nodes
- Edge weights based on co-occurrence data as well as importance of words to documents (Tf-Idf)
- Preference vector consists of query user and words from query document's content

WordTags Recommender

- Simple content-based recommender
- From the co-occurrence matrix of documents and tags, we learn co-occurrence relationships between words and tags

$$weight(w_l, t_k) = \sum_{d_j \in Posts(w_l, t_k)} TfIdf(w_l, d_j)$$

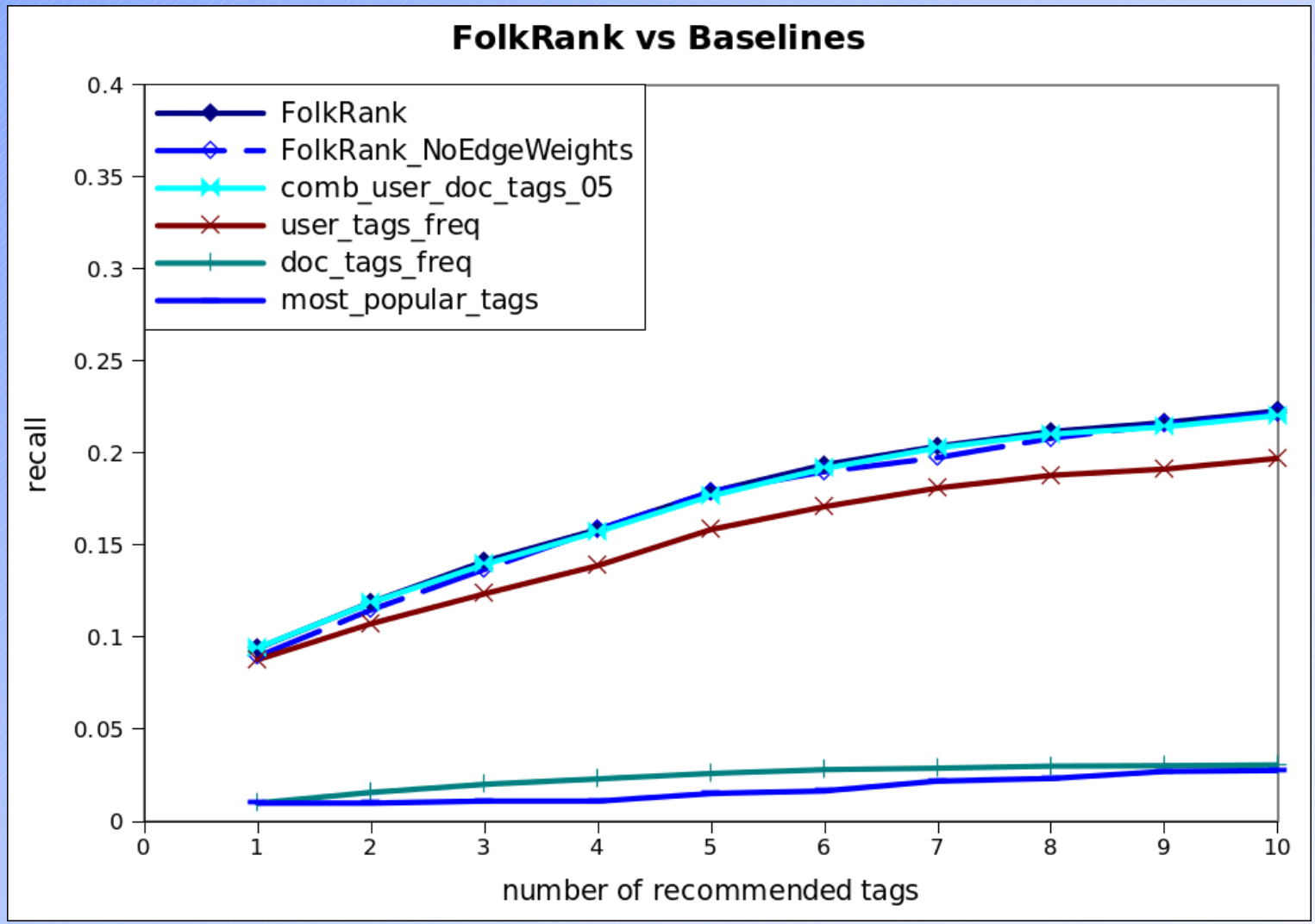
- To recommend tags for a query document d_q we calculate tag scores by

$$score(d_q, t) = \sum_{w_l \in d_q} (TfIdf(w_l, d_q) * weight(w_l, t))$$

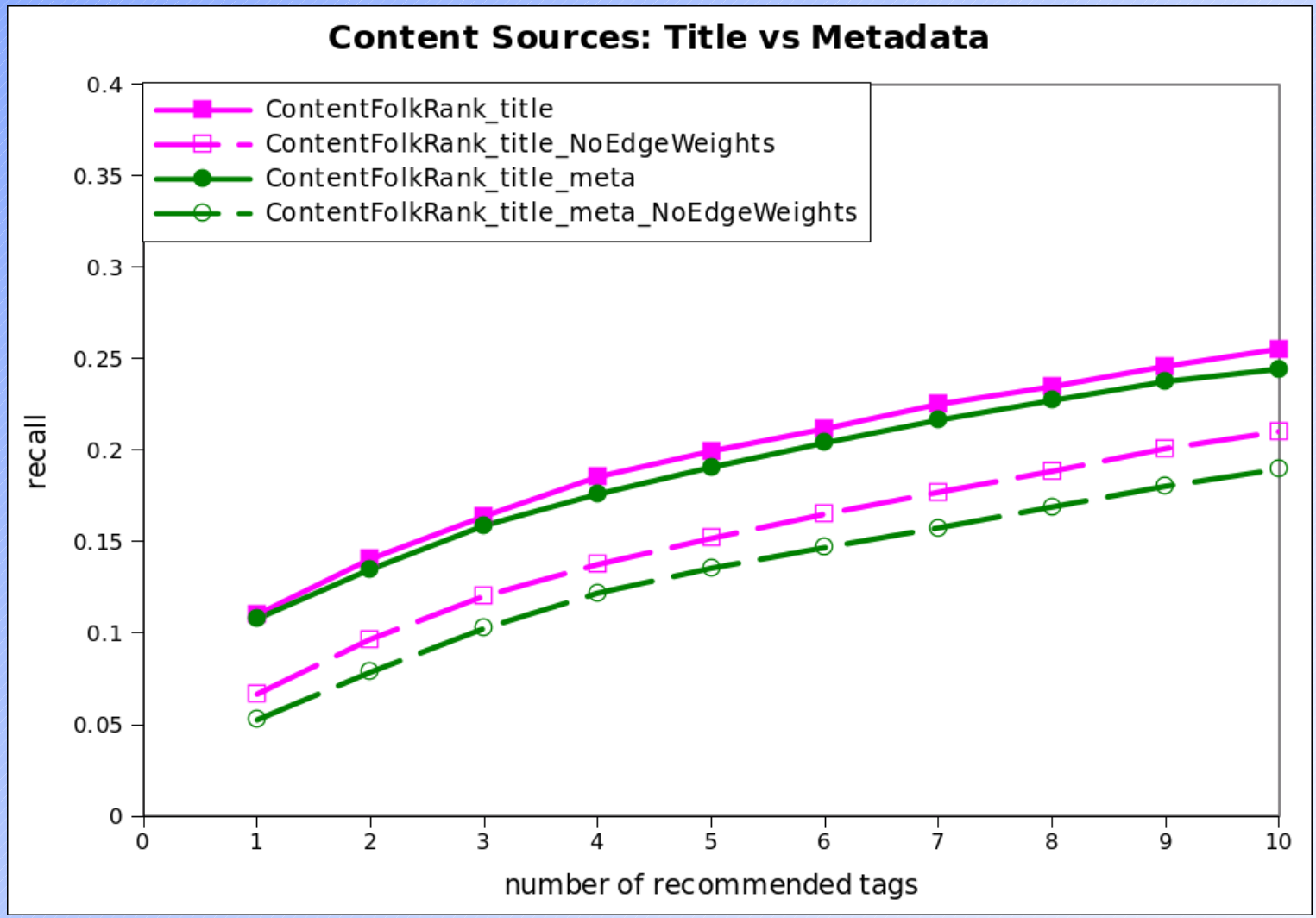
Experimental Setup

- Fixed size N of tag recommendation set
- Evaluation Metric: Recall@N
- BibSonomy Dataset

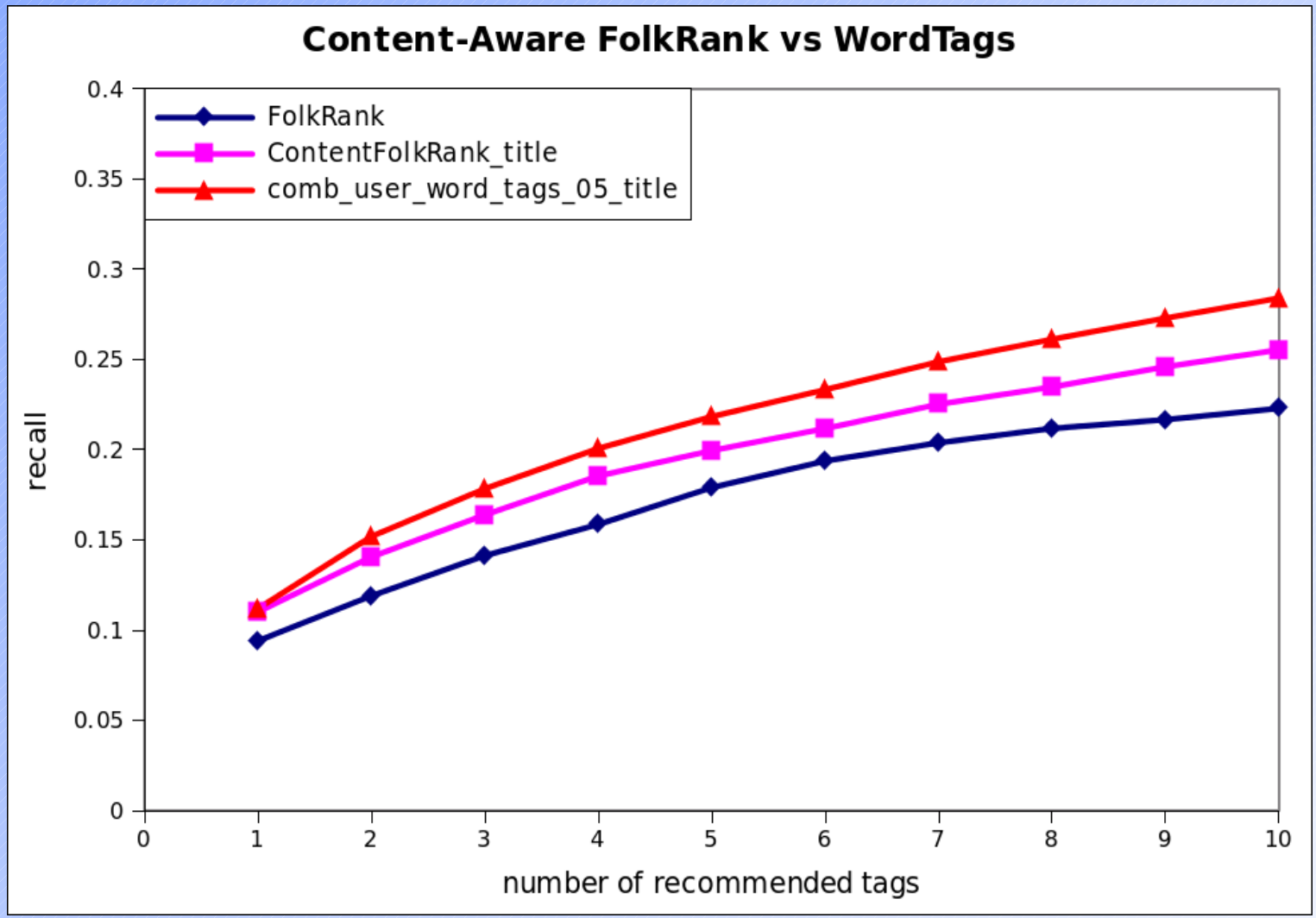
Evaluation Results



Evaluation Results



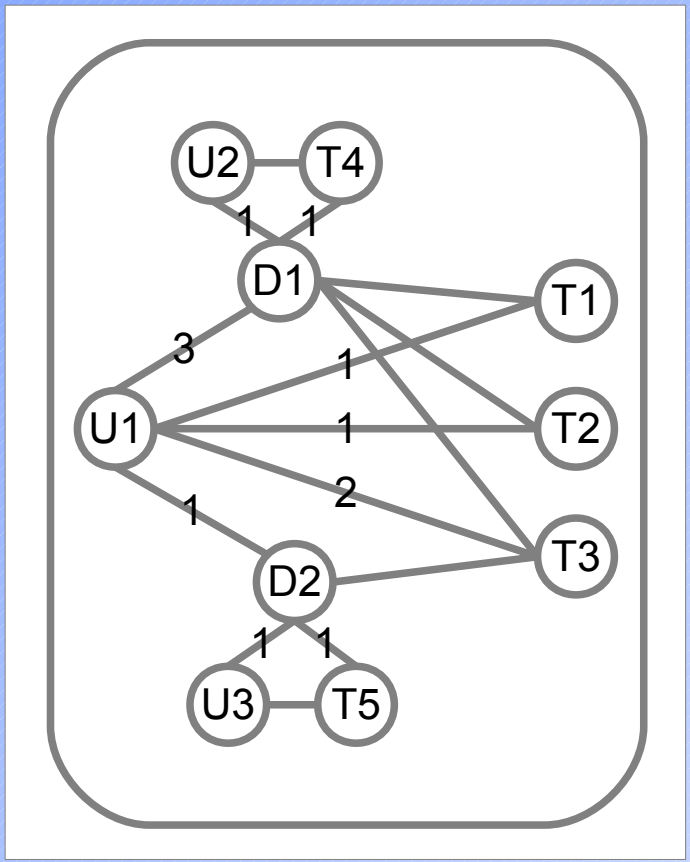
Evaluation Results



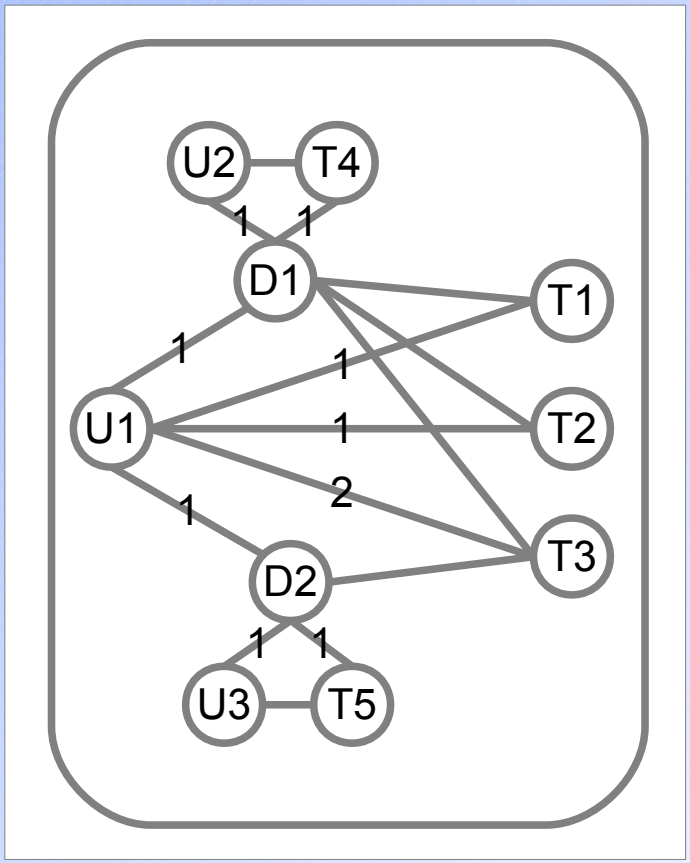
Conclusions

- Content is important and improves recommendation results
- For content-based approaches it is advantageous to include a content-based word importance measure such as Tf-Idf
- Simpler recommender WordTags + UserTags outperforms ContentFolkRank
- UserTags + DocTags performs equally well to FolkRank
- An optimisation of the weighting schemes of FolkRank and ContentFolkRank is worth investigating

Analysis of FolkRank Edge Weights

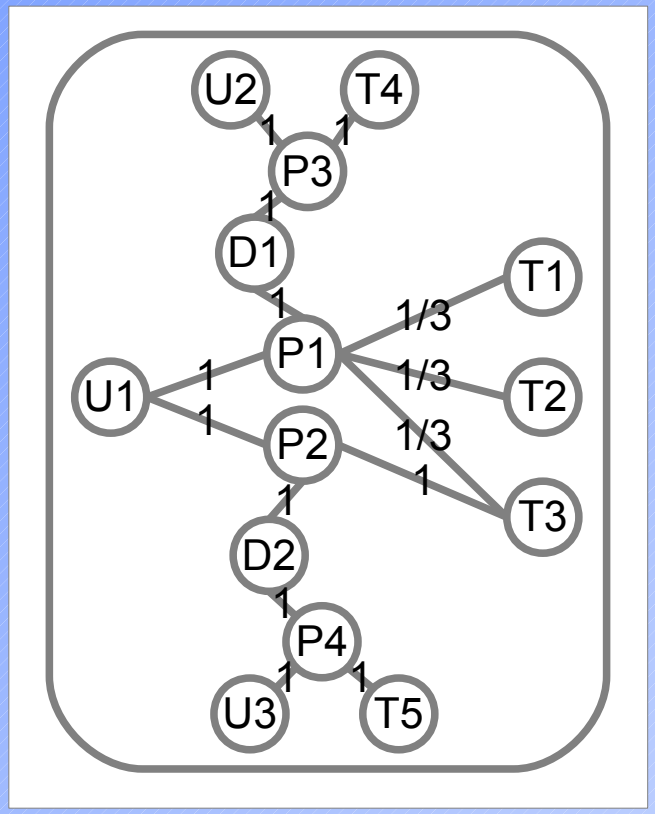


FolkRank

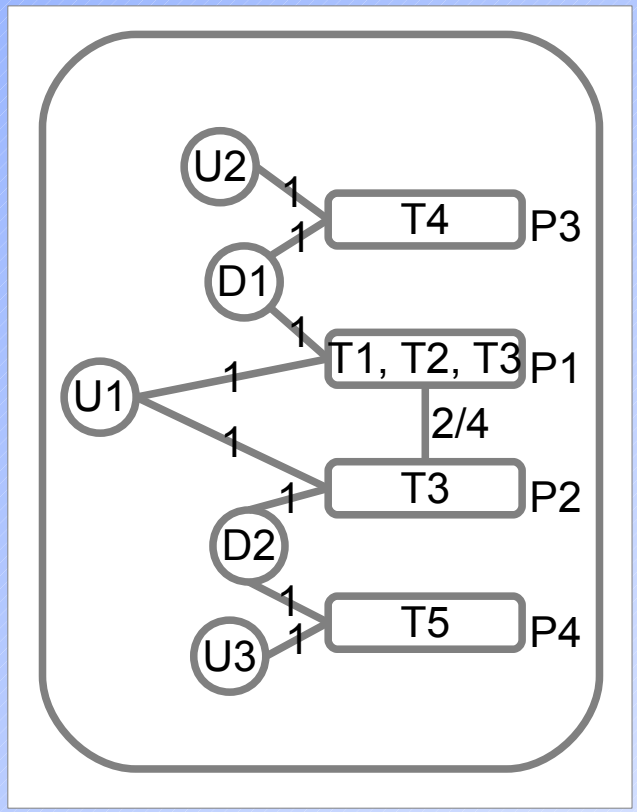


FolkRank2

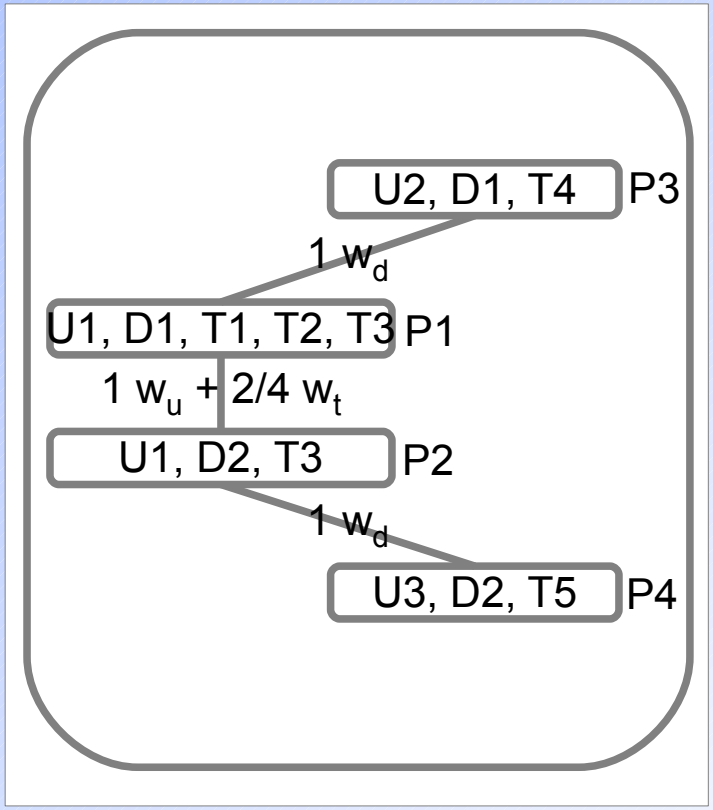
PostRank



PostRank

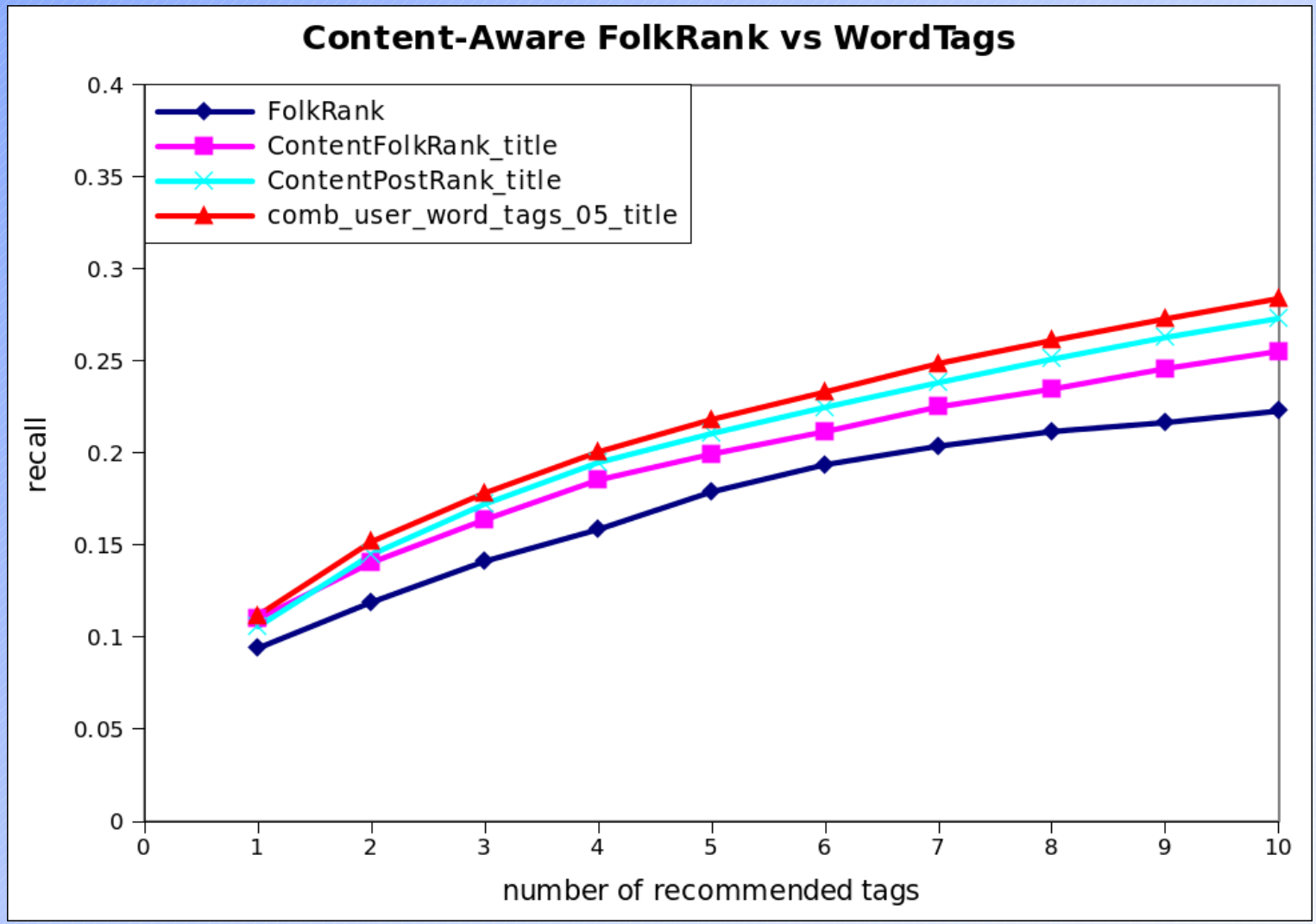


PostRank2



PostRank3

First PostRank Results



Future Work

- Further investigate FolkRank edge weighting scheme
- Investigate issues in FolkRank weight spreading due to the indirected graph: *Swash-back* and *Triangle Spreading*
- Evaluate on CiteULike and Delicious datasets
- Analyse the inherent biases in different sampling/ crawling techniques that are widely used to obtain evaluation datasets

Thanks!

Questions?