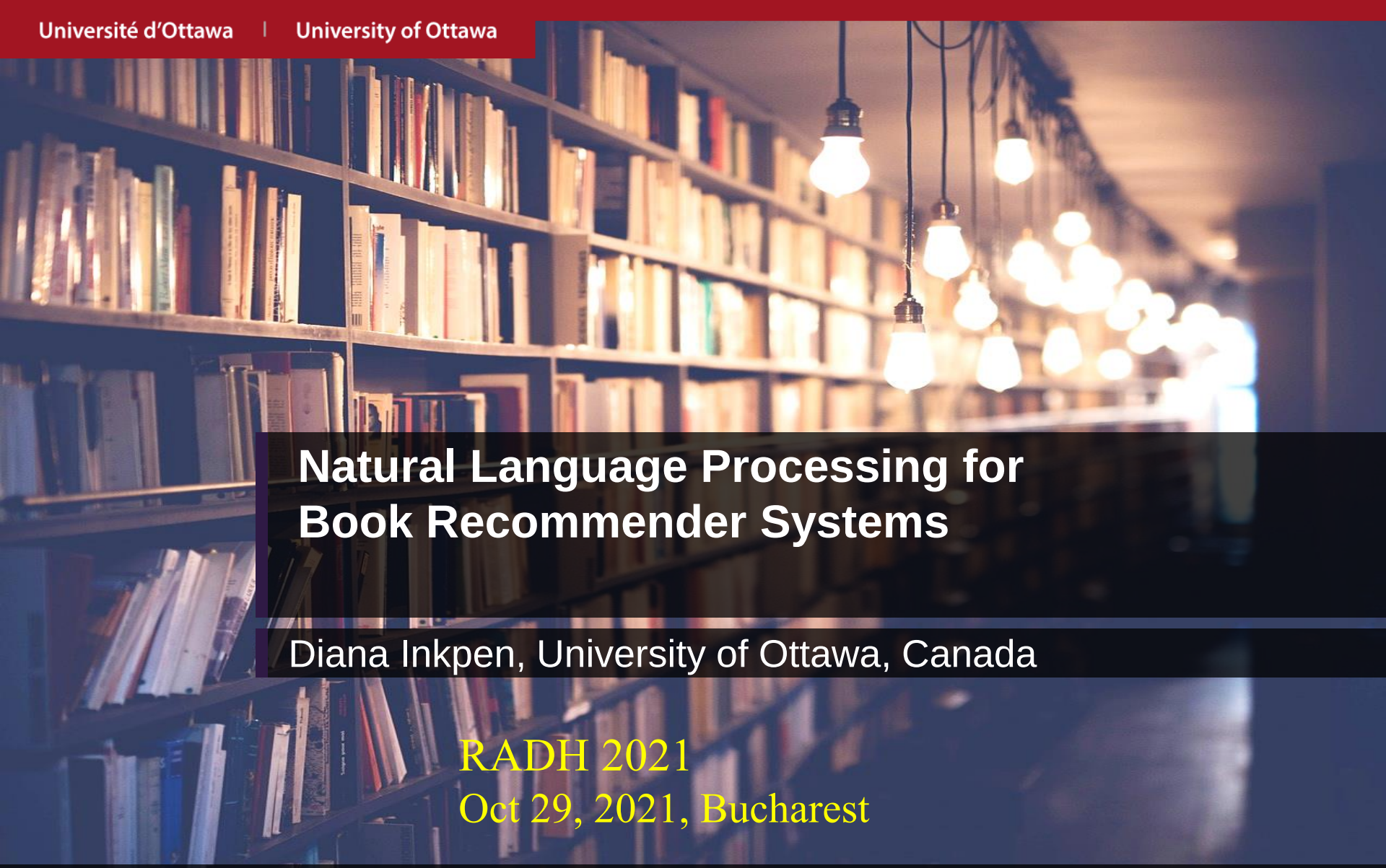


Natural Language Processing for Book Recommender Systems

Diana Inkpen, University of Ottawa, Canada

RADH 2021
Oct 29, 2021, Bucharest

Research area

Artificial Intelligence:

Natural Language Processing:

Automatic text classification and information extraction from various kinds of texts

Social media texts:

- Classifying posts by topic, opinion, emotion
- Cyberbullying detection
- Detecting sings of depression and suicide ideation

Outline

- Introduction to book recommender systems (RS)
- Our book recommender systems
  - Authorship-based RS
  - RS based on linguistic features
  - Topic model-based RS
- Conclusion and future work

Credit: Based on work with my former PhD student Haifa Alharti

Recommender systems (RSs)

Recommender systems predict the items that a user is likely to be interested in.

Examples:

- movie recommenders
- generate playlists of music for streaming services
- product recommenders for online stores
- content recommenders for social media platforms

Collaborative filtering

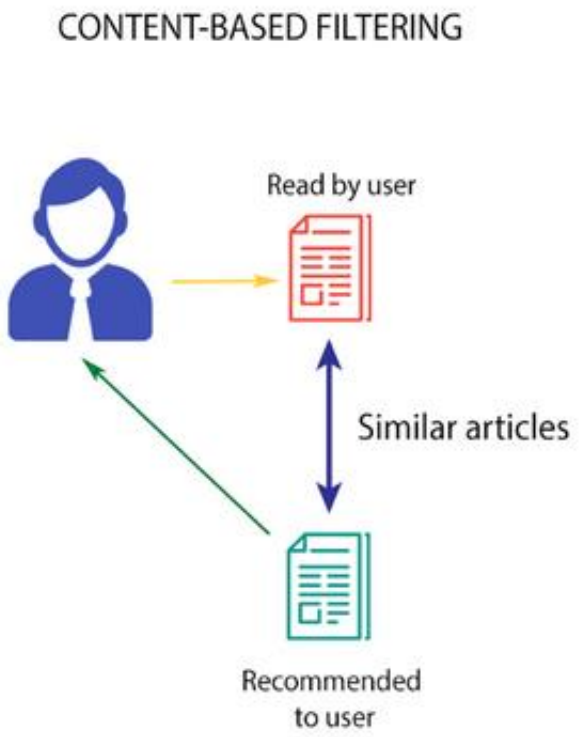
- Technique that can filter out items that a user might like on the basis of reactions by similar users.
- Finding a small set of users with tastes similar to a particular user.
- **User-item matrix with ratings**

		Items					
		1	*2*	...	*i*	...	*m*
Users	*1*	5	3		1	2	
	2		2				4
	:			5			
	u	3	4		2	1	
	:					4	
	n			3	2		

- Methods
  - Memory-based: uses all the data all the time to make predictions.
  - Model based: use the data to learn/train a model which can later be used to make predictions.

Content-based book recommender systems

Recommender systems predict the items (books) that a user is likely to be interested in.



We propose:

- Content based systems that use the texts of books
- A recommendation component that complements them and addresses the new user issue

Motivation

Importance of reading and book RSs

- RSs have advantages for both sellers and consumers.
- Book RSs are useful in libraries, schools, and e-learning portals.
- The practice of reading is declining while many studies proved it is beneficial

The use of book textual content

- Books have distinctive features (e.g., reader advisory's appeal factors)
- Copyrights issues! many initiatives encourage us to follow this direction (e.g., Google Books and BookLamp)

Motivation

User cold start

- Content-based RSs cannot generate personalized recommendations for new users with no rating history
- Signup process (i.e., filling questionnaires) can be long

Problem Statement

Book-recommendations based on textual content

- The use of full texts can lead to better understanding of user reading preferences and to high-quality recommendations.
- Exploit explicit ratings to recommend English literary books in ranking recommendation scenario

Topic modeling in book RSs for new users

- Book RSs should provide new users with quality suggestions, without requiring them to fill long forms
- The availability of user-generated texts on social media provides a chance for RSs to learn information voluntarily shared by users.

Related work

- Only a few took the actual text of the books into account: Vaz et al. (2012a), Zhang and Chow (2015) and Givon and Lavrenko (2009)
- We are not aware of any book RSs that use social media rather than book-cataloguing websites.
- Recommendations based on user Twitter accounts is investigated in news and movie recommendations such as Nair et al. (2016)

First approach:

USING AUTHORSHIP IDENTIFICATION FOR LITERARY BOOK RECOMMENDATIONS

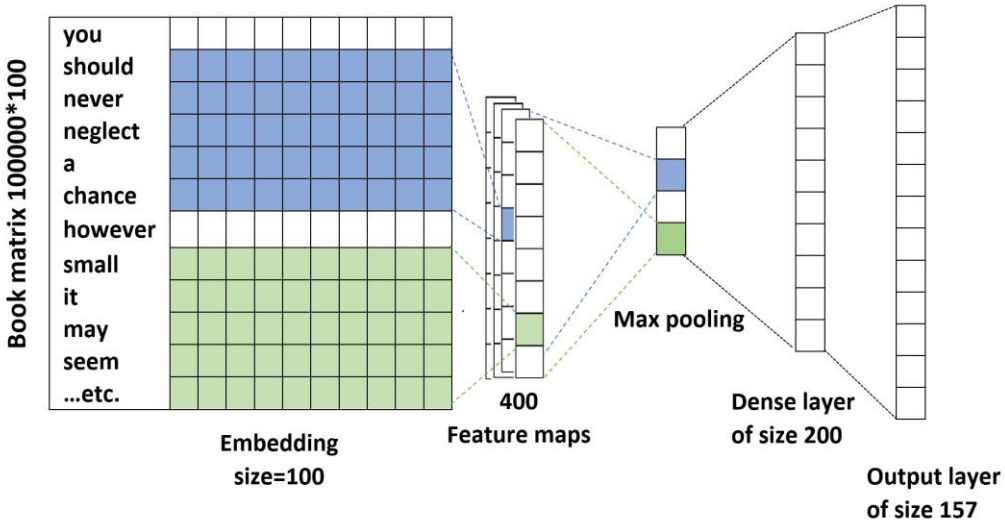
The authorship-based recommender system (AuthId RS)

1- Authorship identification:

- Given the text of a book as input, the AuthId classifier predicts its author, and once it has achieved good accuracy, we extract AuthId book features from the last hidden layer.

2. Book recommendation

- Support Vector Regression (SVR) is trained over book AuthId features associated with the target user ratings to generate a list of ranked books.



CNN for authorship identification

Data

- **Dataset and preprocessing**
  - Litrec Dataset (Vaz et al., 2012c)
  - User ratings from Goodreads: 3-5 (like), 1-2 (dislike), 0 (discarded)
  - Filtered out any user with fewer than 10 books
  - Filtered out any author with fewer than 3 books
  - 351 users, 1010 unique books (from Project Gutenberg) with 157 distinct authors
  - Included the first 100,000 words of each book

Evaluation

The Experimental Setting

- Top-k recommendation scenario
- Similar to many related projects, we set k to 10.
- The system learns from the books in the training set and ranks the remaining books in the dataset
- Three-fold cross-validation is adopted per user, and the results are averaged.

Metrics

$$P@k = \frac{\#\ relevant\ books\ in\ top\ k\ recommended}{k}$$

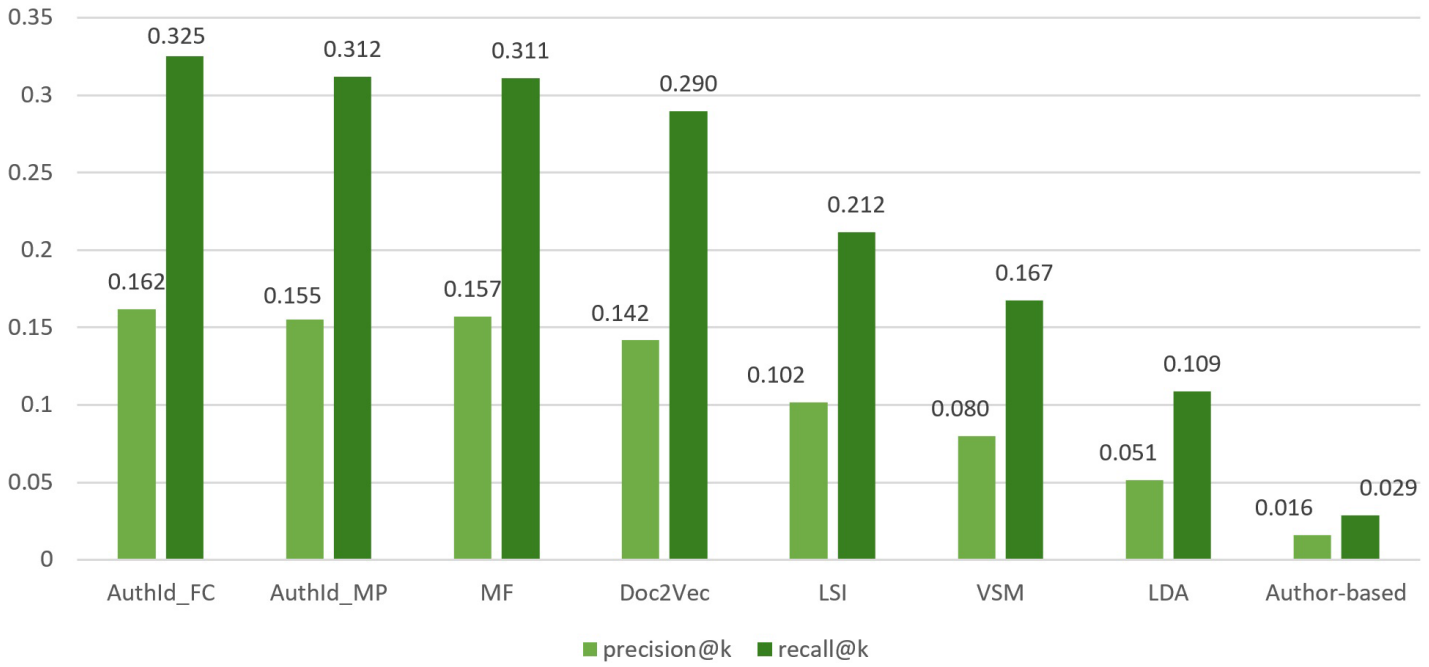
$$R@k = \frac{\#\ relevant\ books\ in\ top\ k\ recommended}{\#\ of\ relevant\ books}$$

Baselines

- ***Content-based* baselines*:***
  - Used the same 100,000 words of each book
  - **Vector space model (VSM)**
  - **LDA and LSI**:
    - 10, 50, 100 and 200 topics
  - **Doc2vec**:
    - 100 vs. 300 dimensions and 10 vs. 5 window size
  - **Plain author-based RS**
    - Support Vector Regression (SVR)
    - instead of book AuthId representations, author ids are used.
- **Collaborative filtering (Matrix Factorization)**
  - Recommends books read by similar users
  - 8, 16, 32 and 64 latent factors.

Top-k recommendation accuracy

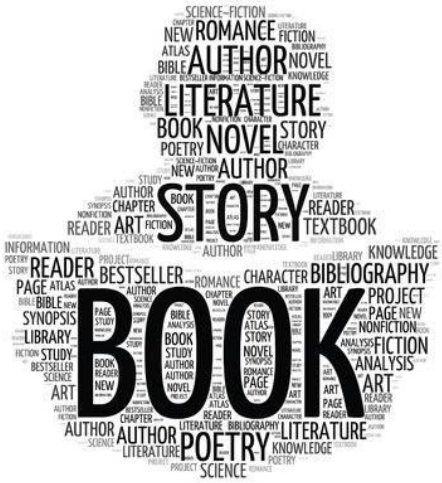
In 17 test cases, the system could rank in the top ten list at least one relevant book that was written by an author who did not appear in the training set.



P@10 and R@10 generated by our system and the baselines.

Analysis

- The shorter the text used in the source model the lower recommendation performance
- Many relevant books ranked in the top ten list are retrieved from a large collection of books written by the same authors.
- Similar book representations are annotated similarly by experts on Novelist



Second approach:

BOOK RECOMMENDATIONS BASED ON LINGUISTIC FEATURES

Linguistic Features Incorporated in a Literary Book Recommender System

Features selection:

- 120 features to describe content and style of each book

Recommendation procedure

- Find the most similar book representation using K-nearest neighbors (KNN)
- Used Extremely Randomized Trees (ET) regressor to predict the ratings and rank books

Categories of features included in our study

Category	Feature Names
Lexical	*Token-based:* average length of paragraph, sentence and words; average number of commas per sentence; average variance in paragraph length and in sentence length; length of book; words>6-letters; dictionary words; average syllables per word Lexical density (type-token ratio) *Word frequencies:* percent of latinate words; function words; affect words; social words; cognitive processes; perpetual processes; biological processes; core drives and needs; time orientation; relativity; personal concerns[2]
Character-based	Numbers; all punctuations
Syntactic	Percentage of adjectives, adverbs, nouns, and verbs; comparatives, interrogatives; quantifiers
Fiction-based	Percentage of dialog text; the number of dialogs; the average length of dialog texts; the number of unique fictional characters; the number of times fictional characters are mentioned; the number of unique places; the number of times places are mentioned; Percentage of dialogue by female characters; Percentage of characters which are female
Six-styles	Literary; abstract; objective; colloquial; concrete; subjective

Example of results

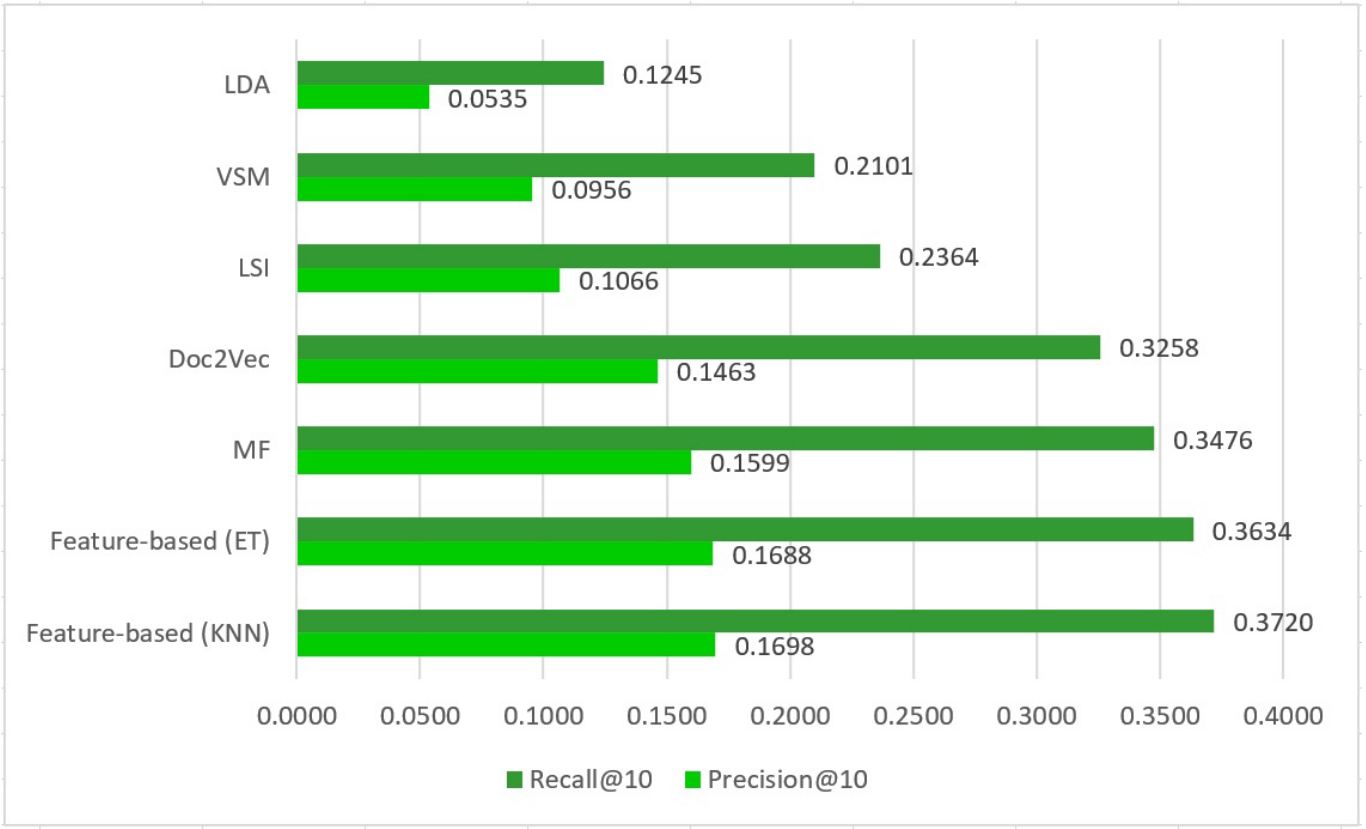
Book name (similarity)	Description on NoveList
1- Madame de Treymes by Edith Wharton (1) 2- The Descent of Man and Other Stories by Edith Wharton (0.72) 3- The Glimpses of the Moon by Edith Wharton (0.71)	*Genre:* **Love stories**, Modern classics; *Character*: **Complex**; *Storyline*: **Character-driven** ; *Writing Style*: Descriptive, Lush, Richly detailed; *Time Period:* 1920s; *Location*: Europe – Social life and customs – 20th century; *Subject headings*: Social status, Freeloaders, Americans in Europe, Honeymoons, **Married people**, Misunderstanding, **Men/women relations**
4- Our Friend the Charlatan by George Gissing (0.71) 5- Denzil Quarrier by George Gissing (0.698)	*Genre*: Psychological fiction, Domestic fiction, **Love stories**, Historical fiction, Satirical fiction; *Time Period*: 19th century; *Subject headings*: Women – Employment - England, Middle class women - England, Women England, Single women - England, Sisters - England, Married women - England, Self-discovery in women, Authors, Fiction writing, Married people; *Location*: London, England
6- A Daughter of To-Day by Sara Jeannette Duncan (0.697)	*Genre*: Canadian fiction; *Time Period*: 20th century; *Subject headings*: Brothers and sisters, Politicians, Romantic love, Clergy, Political science, **Men/women relations** ; *Location*: Ontario
7- In the Year of Jubilee by Gissing (0.69)	See above
8- The Tragic Muse by Henry James (0.687) 9- Sir Dominick Ferrand by Henry James (0.682)	*Genre*: Classics, Psychological fiction; *Character*: **Complex**, Flawed; *Storyline*: **Character-driven**; *Pace*: Leisurely paced; *Tone*: Melancholy, Reflective, Thought-provoking; *Writing Style*: Stylistically complex
10- The Tragic Bride by Francis Brett Young (0.678)	Genre: Domestic fiction *Subject headings*: Young women, Quests, Love, **Men/women relations**, Marriage, Remarriage, Self-discovery in women, Happiness in women; *Location*: Midlands, England

Top ten similar books to Madame de Treymes by Edith Wharton

Evaluation

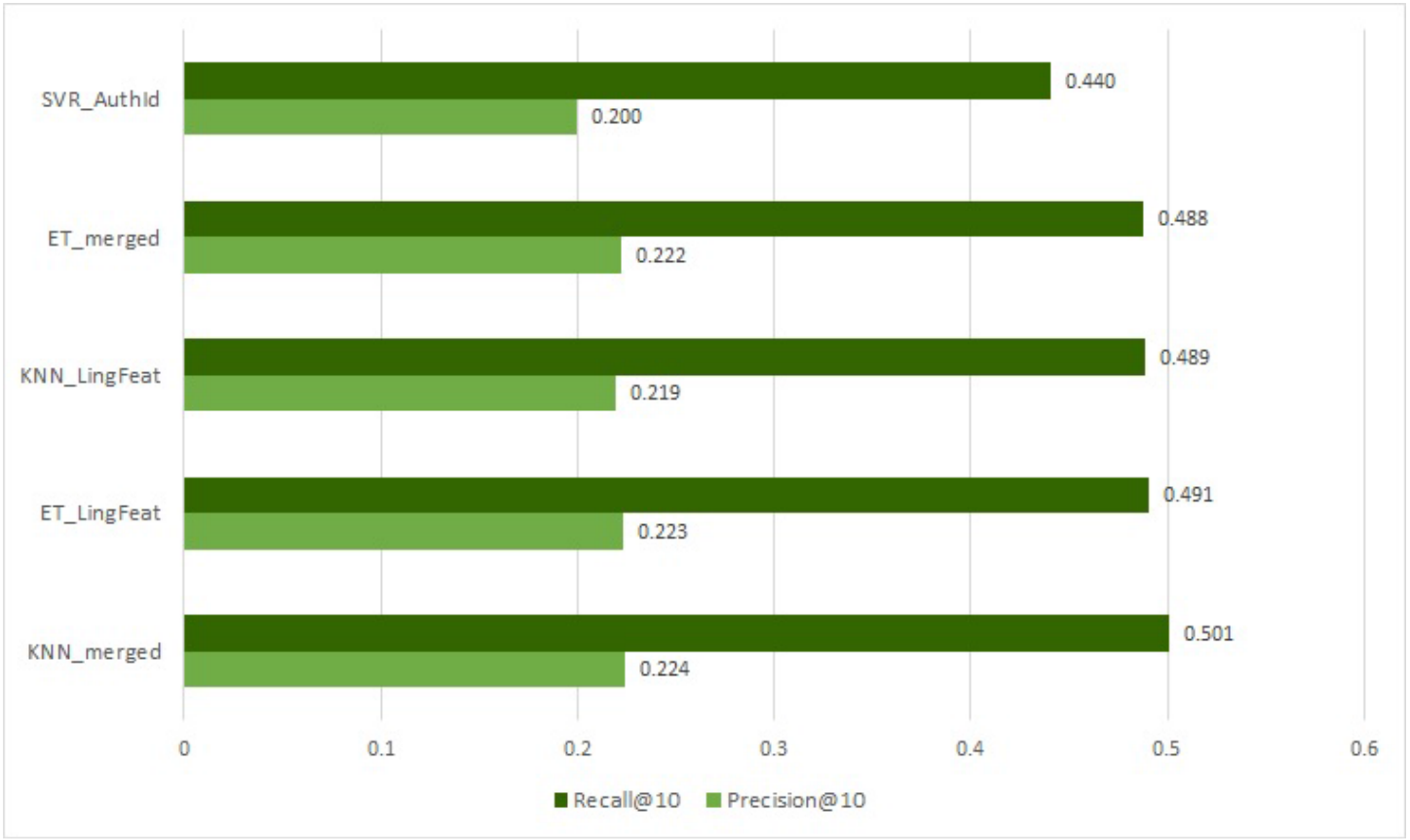
- **Dataset and preprocessing**
  - Litrec Dataset (Vaz et al., 2012c)
  - Filtered out 71 books without fictional characters
  - 367 users, 1,050 unique books and 410 different writers
- **The Experimental Setting**
- **Metrics** (P@10 and R@10)
- **Baselines**

Top-k recommendation accuracy



Recommendation accuracy of our approach against the baselines

Comparison of linguistic features vs. AuthId book representations



Analysis

- Highlighted the features with highest/lowest averaged importance values generated by 730 accurate ET models
- Generated recommendations based on the most important features and KNN showed best performance when top 80 features are used
- Similar book representations are annotated similarly by experts on Novelist

Third approach:

TOPIC MODEL-BASED (TMB) SYSTEM FOR COLD USER ISSUE

TMB for new users

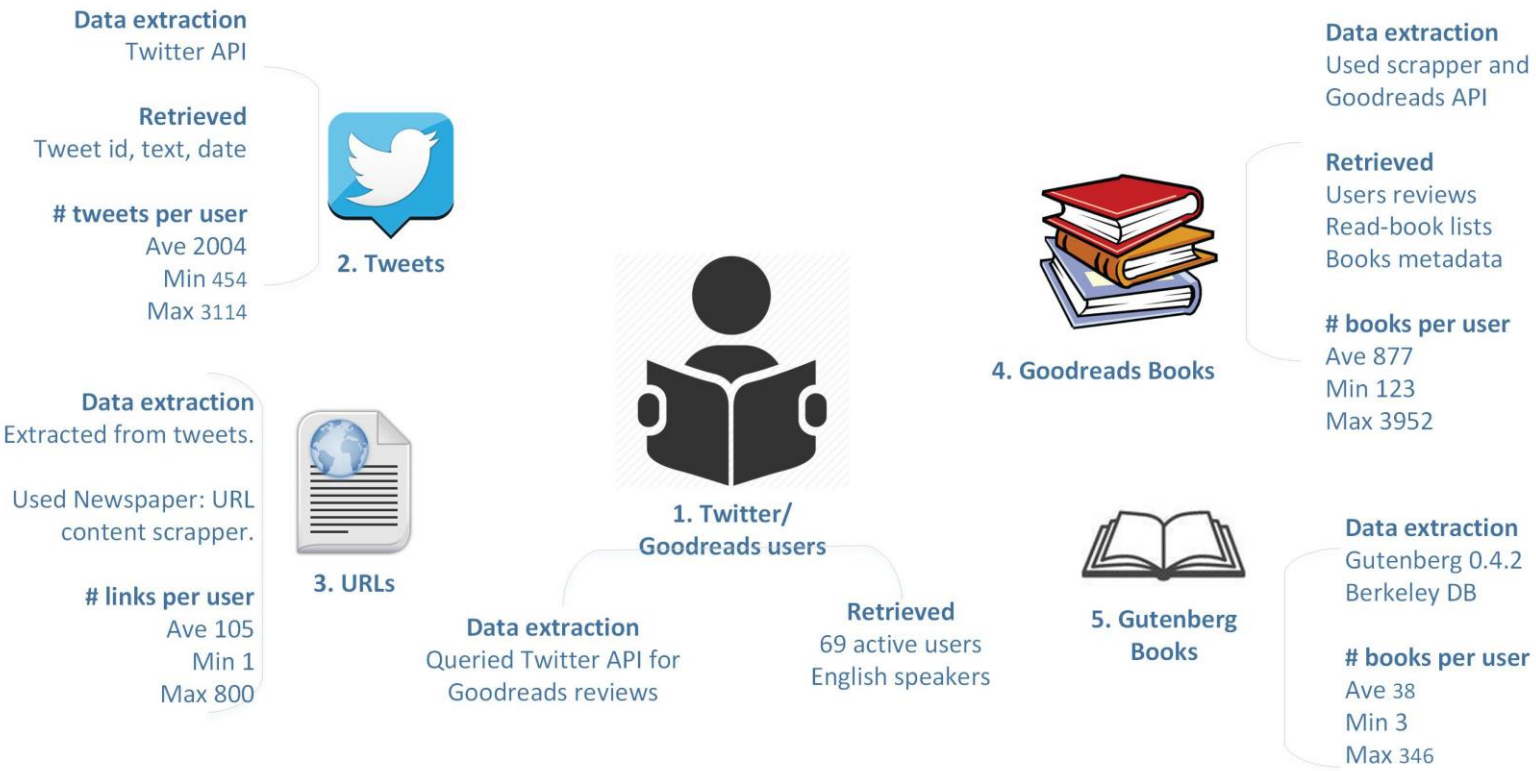
Book and user profiles

- **User profiles (UP)**
  - Vectors of terms extracted from tweets and/or links
  - Top tf-idf terms (a user timeline as a document)
- **Book profiles (BP)**
  - Vectors of terms extracted from book descriptions
- **Topic embeddings**
  - Terms are mapped to word2vec pre-trained word embeddings

Recommendation procedure

- Once enriched with word embeddings, averaged embeddings of words in UP and BP are averaged
- Cosine similarity is found between UP and BP
- Evaluation: showed similar performance to a meta-data based system.

Evaluation: Data collection



Dataset collection and statistics.

Data preprocessing

- Tokenization and POS tagging
- Noun-based user and book profiles
- Removal of useless information from user profiles:
  - Web-related terms
  - Tweets with Goodreads links
  - 100 most common English nouns
  - Words of three letters or less
  - Misspelled hashtags
  - Words with low *idf* values

The Experimental Setting

- **Baselines:** random system and metadata-based CB
- A time threshold is set to avoid overlap between learning and prediction times
- **Leave-one-out evaluation:** retrieve one test book out of 1000 random books irrelevant to all users. Then select 1 relevant book per user and see if it can be retrieved at high rank (*f*)
- **Metrics:** hit rate and average reciprocal hit-rank
- Five trials per user were conducted and the results are averaged.

$$HR = \frac{\#hits}{\#users}$$

$$ARHR = \frac{1}{\#users} \sum_{i=1}^{\#hits} \frac{1}{f_i}$$

Results

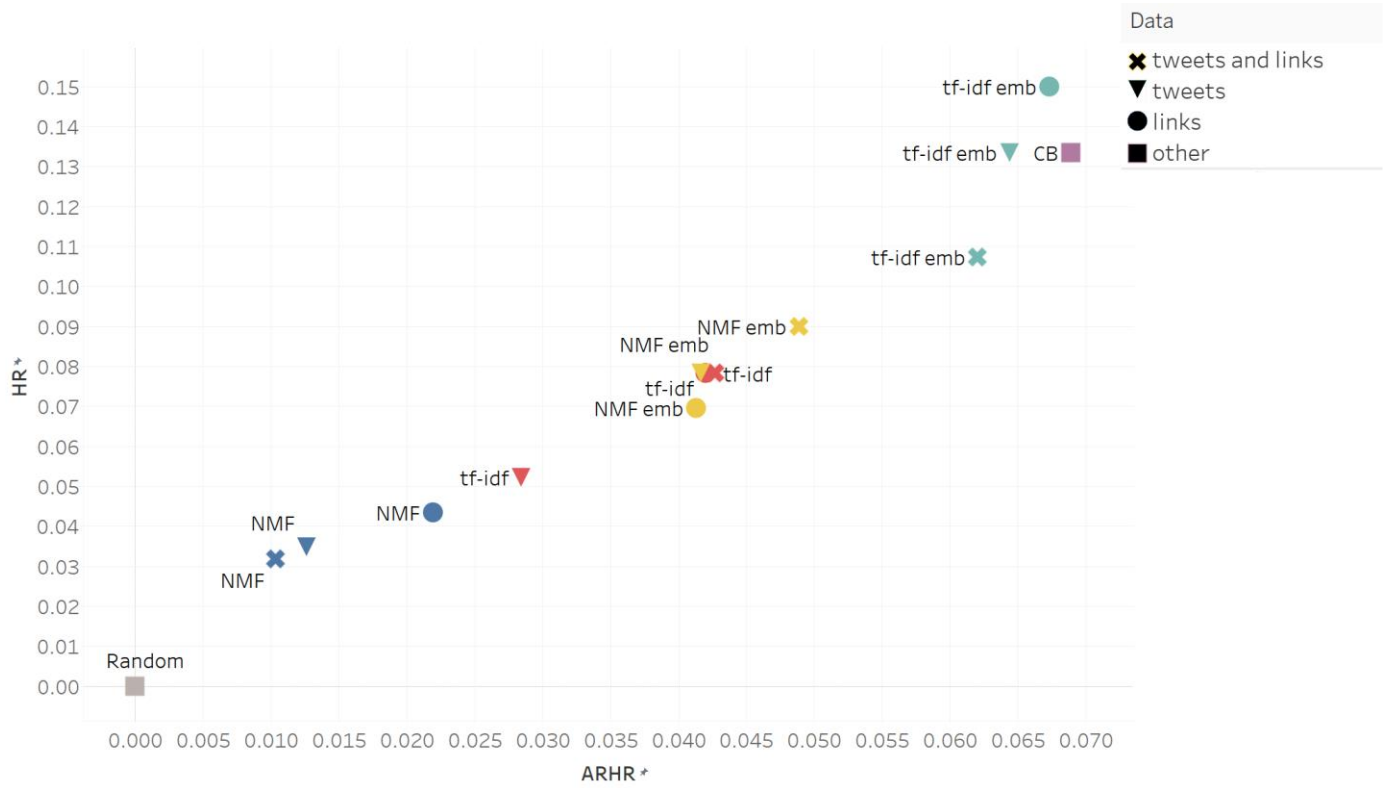


Figure 5.2: The comparison of TMB approaches, CB and the random system.

Summary

1. Authorship-based RS:
   - The first to transfer information learned by an author-identification model to book recommendation.
   - gives higher recommendation accuracy than CF and CB baselines.
2. RS based on linguistic features:
   - The first to include and analyze a large number of linguistic features for book recommendation.
   - It outperforms several baselines and provides variable importance values.
3. Topic model-based recommendations
   - The first to extract user-discussed subjects from social media and map them to books.
   - Collected a dataset that contains users' Twitter accounts and Goodreads's book preferences.
   - Showed similar performance to a meta-data based system.

Limitation

- Overspecialization
- Lack of content
- The source model in AuthId RS needs to be retrained when new books/authors are added
- TMB is a useful when users are active on social media and willing to share their data

Future work

- Multi-task classifier to predict (author, genre) or (author, user ratings)
- Different architectures for authId source model
- Recommendation of books from various times, genres, and languages
- Recommend books for non-readers

Published Papers

1. H. Alharthi, D. Inkpen, and S. Szpakowicz. A survey of book recommender systems. Journal of Intelligent Information Systems, pages 1–22, 9 2017a. ISSN 0925-9902. doi:10:1007/s10844-017-0489-9. URL http:https://doi:org/10:1007/s10844-017-0489-9
2. H. Alharthi, D. Inkpen, and S. Szpakowicz. Unsupervised topic modelling in a book recommender system for new users. In SIGIR 2017 Workshop on eCommerce (ECOM17), 2017b. ISBN 978-1-4503-5022-8. doi: 10:1145/3077136:3084367. URL http://doi:acm:org/10:1145/3077136:3084367
3. H. Alharthi, D. Inkpen, and S. Szpakowicz. Authorship identification for literary book recommendations. In Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, pages 390–400, 2018. URL https://aclanthology:info/papers/C18-1033/c18-1033
4. H. Alharthi and D. Inkpen. Study of Linguistic Features Incorporated in a Literary Book Recommender System. In Proceedings of the 34th ACM/SIGAPP Symposium On Applied Computing (SAC '19), New York, NY, USA, April 2019. ACM. doi: https://doi:org/10:1145/3297280:3297382

Thank you! Mulțumesc!