Методы анализа научных публикаций и их взаимосвязей

Акопян Микаэла Тиграновна

научный руководитель: к.т.н. Григорьева Мария Александровна

Что такое рекомендательные системы?

Рекомендательная система — это класс алгоритмов, которые анализируют большие массивы данных и выявляют закономерности, позволяющие делать автоматизированные выводы и предсказания.

В научной аналитике методы, лежащие в основе рекомендательных систем, используются для:

- анализа содержания научных статей,
- изучения связей между авторами и публикациями,
- выявления структур научных областей,
- определения факторов, влияющих на цитируемость.

Необходимость автоматизации анализа данных в науке

- Стремительный рост числа научных статей
- Информационная перегрузка
- Сложность анализа вручную
- Неравномерная видимость статей
- Неочевидные факторы повышения цитируемости

Основные методы анализа связанности данных в научной среде

- Content-based методы
  (анализ текста, извлечение признаков статьи)
- Distributed representations
  (векторные представления слов и документов)
- Transformer-based модели
  (контекстные embeddings научных статей)
- Collaborative filtering
  (анализ совместных действий и структур взаимодействий)
- Graph-based методы
  (анализ сетей цитирований, соавторств и ключевых слов)

Bag-of-words подходы (мешок слов)

1. TF-IDF (Karen Spärck Jones и Stephen Robertson, 1972)

TF = Количество вхождений термина t в документ d / общее количество слов в документе

IDF = log(M/(1 + df(t)), M – общее количество документов, df(t) –количество документов, содержащих термин t.

TF-IDF = TF * IDF

2. BM25 – улучшение TF-IDF (Karen Spärck Jones и Stephen Robertson, 1980-е))

Distributed representations

1. Word2Vec (Google, 2013)

Основные режимы:

CBOW (Continuous Bag of Words) – по окружающим словам модель предсказывает центральное слово.

Skip-gram – по центральному слову модель предсказывает окружающие

2. Doc2vec (Quoc V. Le, Tomas Mikolov, 2014)

Основные режимы:

PV-DM: модель по (document vector + context words) предсказывает следующее слово в тексте

PV-DBOW: модель по вектору документа учится предсказывать слова, которые в нём встречаются (без контекста)

Transformer-based подходы

1. SciBERT (Iz Beltagy, Kyle Lo, Arman Cohan, 2019)

Создана на основе BERT. SciBERT обучен на ≈1,14 млн научных статей (Semantic Scholar). Около 80% — CS.

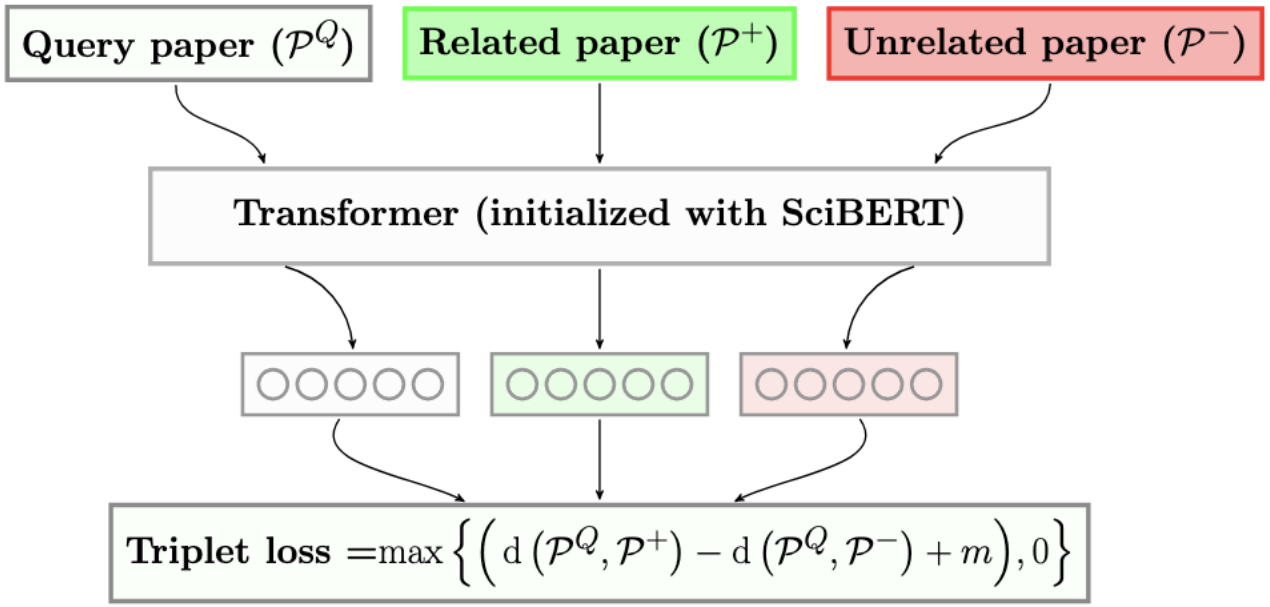
SciBERT не использует словарь BERT. Она использует специальный научный словарь SciVocab.

CS	NER	SciERC (Luan et al., 2018)	64.20	63.58	65.24	65.77	**67.57**
	REL	SciERC (Luan et al., 2018)	n/a	72.74	78.71	75.25	**79.97**
	CLS	ACL-ARC (Jurgens et al., 2018)	67.9	62.04	63.91	60.74	**70.98**

Transformer-based подходы

2. SPECTER (Arman Cohan, Sergey Feldman, Iz Beltagy, 2020)

Расширение SciBERT. Обучена на тексте + графе цитирований.



Task →	Classification		User activity prediction				Citation prediction				Recomm.		Avg.
Subtask →	MAG	MeSH	Co-View		Co-Read		Cite		Co-Cite				
Model ↓ / Metric →	F1	F1	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	nD$\hat{C}$G	P@$\hat{1}$	
Random	4.8	9.4	25.2	51.6	25.6	51.9	25.1	51.5	24.9	51.4	51.3	16.8	32.5
Doc2vec (2014)	66.2	69.2	67.8	82.9	64.9	81.6	65.3	82.2	67.1	83.4	51.7	16.9	66.6
Fasttext-sum (2017)	78.1	84.1	76.5	87.9	75.3	87.4	74.6	88.1	77.8	89.6	52.5	18.0	74.1
SIF (2017)	78.4	81.4	79.4	89.4	78.2	88.9	79.4	90.5	80.8	90.9	53.4	19.5	75.9
ELMo (2018)	77.0	75.7	70.3	84.3	67.4	82.6	65.8	82.6	68.5	83.8	52.5	18.2	69.0
Citeomatic (2018)	67.1	75.7	81.1	90.2	80.5	90.2	86.3	94.1	84.4	92.8	52.5	17.3	76.0
SGC (2019a)	76.8	82.7	77.2	88.0	75.7	87.5	**91.6**	**96.2**	84.1	92.5	52.7	18.2	76.9
SciBERT (2019)	79.7	80.7	50.7	73.1	47.7	71.1	48.3	71.7	49.7	72.6	52.1	17.9	59.6
Sent-BERT (2019)	80.5	69.1	68.2	83.3	64.8	81.3	63.5	81.6	66.4	82.8	51.6	17.1	67.5
SPECTER (Ours)	**82.0**	**86.4**	**83.6**	**91.5**	**84.5**	**92.4**	88.3	94.9	**88.1**	**94.8**	**53.9**	**20.0**	**80.0**

Transformer-based подходы

3. SPECTER2 (AllenAI, 2023)

Следующее поколение SPECTER, обученное на более крупном и современном корпусе.
Учитывает структуру графа, текст и метаданные. Даёт ещё более устойчивые и точные embeddings статей.