Personalized recommendation of interesting texts

Michal Kompan
supervisor: prof. Mária Bieliková

Contribution

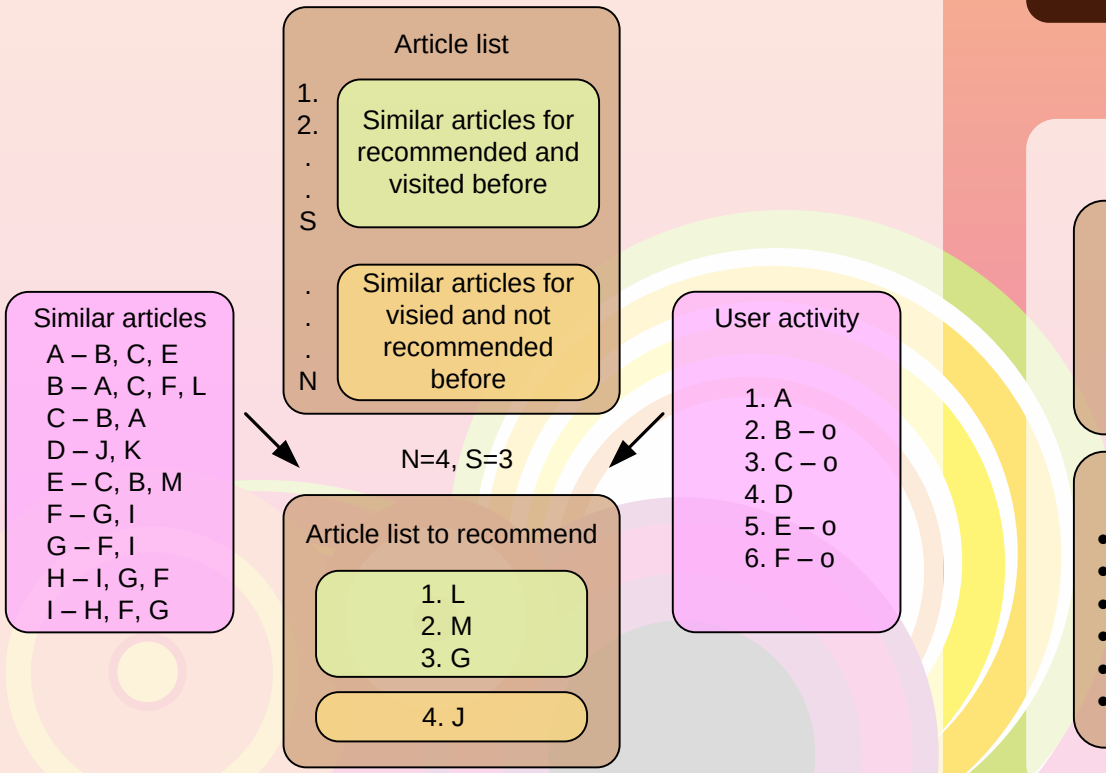
- Short and high representative article vectors
- Real-time content-based recommendation
- Fast similarity computation

Evaluation

- Recommendation : Synthetic tests – SME.SK (3 days)
- Article similarity : SME.SK (Author's choice, manually annotated)

	Train period [h]	Test period [h]	Precision		Recall		F1-Measure	
			Cos.	*Jacc.*	*Cos.*	*Jacc.*	*Cos.*	*Jacc.*
Categories	9	63	43.23	64.05	50.28	36.27	46.49	46.31
	24	48	40.26	63.26	50.94	37.44	44.97	47.04
	33	39	39.73	62.12	51.36	39.92	44.80	48.63
	48	24	38.02	59.91	59.95	40.23	46.53	48.14
Articles	9	63	1.43	1.83	0.84	0.77	1.06	1.08
	24	48	0.76	1.81	0.47	0.80	0.58	1.11
	33	39	0.67	1.68	0.49	0.85	0.57	1.13
	48	24	0.5	1.53	0.64	1.34	0.56	1.43

Recommendation



Article similarity

1. Article preprocessing

- stopwords elimination
- lemmatization
- name and place extraction
- keywords extraction

3. Similarity computation

- cosine similarity
- real-time

$$similarity = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} a_{ji}\, b_{ji}}{\sqrt{\sum_{j=1}^{m}\sum_{i=0}^{n} a_{ji}^2}\sqrt{\sum_{j=1}^{m}\sum_{i=0}^{n} b_{ji}^2}}$$

2. Vector representation

- **Title** – lemmatized words from article title (aprox 5)
- **TF of title words in the content** – normalized
- **Names/Places** – words starting with uppercase and no sentence end before
- **Keywords** – 10 most relevant keywords (TF-IDF)
- **Category** – „tree-based" category vector
- **CLI** – Coleman-Liau readability index