The 18th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems
June 15 - 19, 2020, Lisboa, Portugal

Using Topic Information to Improve Non-Exact Keyword-Based Search for Mobile Applications

Eugénio Ribeiro[1,2], Ricardo Ribeiro[1,3], Fernando Batista[1,3], and João Oliveira[3]

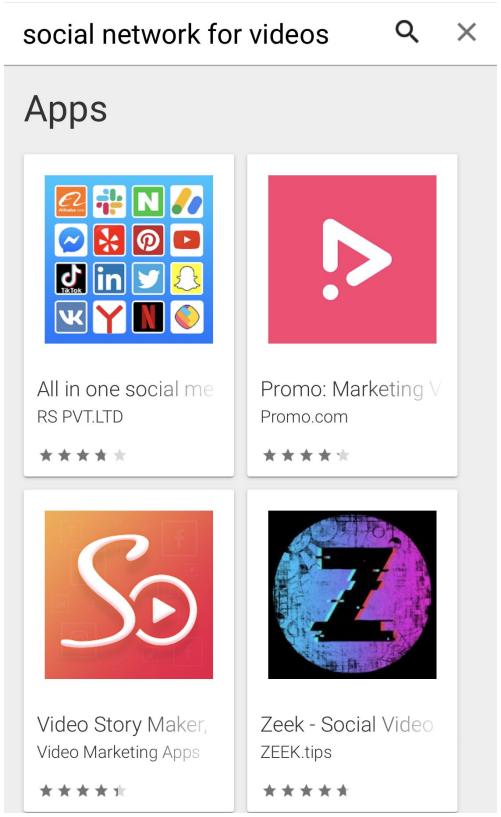
[1] INESC-ID Lisboa, Portugal
[2] Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3] Instituto Universitário de Lisboa (ISCTE-IUL), Portugal

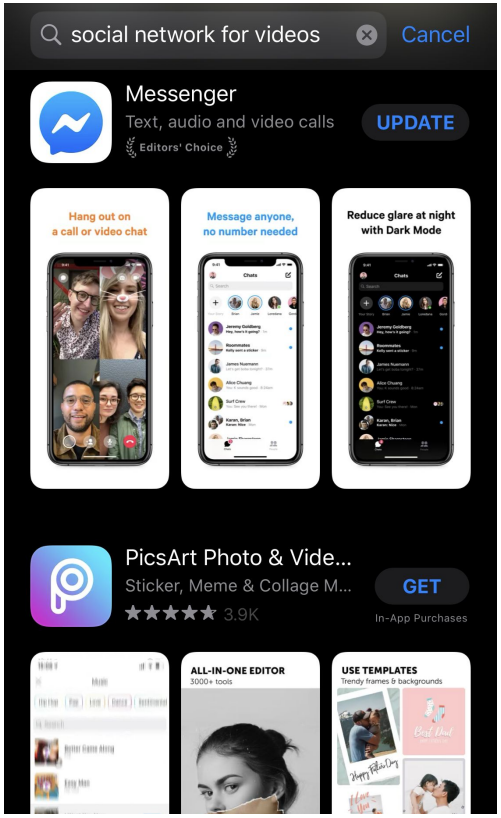
Mobile App Search

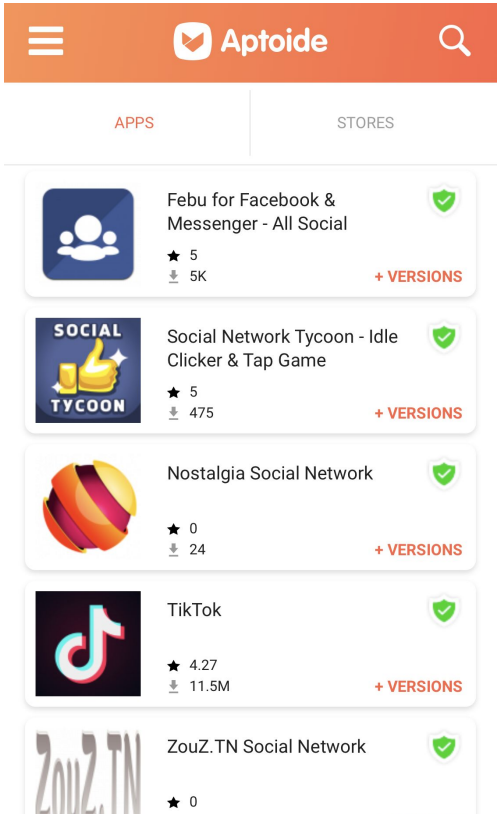
- Mobile apps play an important role in everyday life
- Offer is constantly increasing
  - Significant time searching for new applications
  - Effective search and recommendation engines are required
- Most searches in app stores are exact
  - Users target specific applications
  - Suggested by acquaintances ✔️
  - Found using web search ❌ (Counter-intuitive)

Non-Exact Search

“Social network for videos”







Non-Exact Search

- Search engines are typically keyword-based
  - Unable to semantically interpret queries
- Web search engines have access to more data
  - Application DB vs. WWW
  - Several web pages are dedicated to application comparison
- App store search engines must overcome data problem
  - Semantic interpretation (e.g. using user context)
  - Inference of additional information
  - Bridge the gap between users and developers
  - Increase the match ratio

Topic Information

- Enables matching without using the same words
  - Different words that refer to the same context

- Proved important for information retrieval
  [Yi & Allan 2009]
  - Including search for applications
    [Park et al. 2015][Zhuo et al. 2015]

- Retrieval based on topic distribution similarity

However…

- Most app store search engines:
  - Keyword matching
  - Popularity boosting

Topic Information for Keyword-Based App Search

1. Preprocessing
2. Topic Modeling
3. Topic Keyword Generation
4. Application Retrieval

Description Preprocessing

- Dependency-parsing
  - Sentence splitting
  - Negation identification
- Part-of-speech tagging
  - Enables selection of specific word classes
- Lemmatization
  - Improves match ratio
  - Also performed on queries
- Discarding irrelevant terms
  - Common stopwords
  - High frequency terms
  - Low frequency terms

Topic Modeling

- Latent Dirichlet Allocation (LDA)
  [Blei et al. 2003]

- Terms:
  - Nouns
  - Adjectives (+ Negation)
  - Non-auxiliary Verbs (+ Negation)

- Description-level model:
  - Generic topics
  - May infer non-existing relations

- Sentence-level model:
  - More constrained topics
  - Directly related terms

Topic Keyword Generation

- Common approach:
  - Top *n* terms
  - May include irrelevant terms for one topic
  - May discard important terms for other topic

- Our approach:
  - Term distribution ~ Negative exponential function
  - Identify the inflexion point (i)
  - Select the terms that appear before i in the distribution

Application Retrieval

- Applications represented by 3 fields in the DB
- Application description (a_d)
  - Preprocessed textual description
- Description-level topic keywords (a_{dt})
  - Obtain description-level topic distribution of a_d
  - Merge topic keywords of the relevant topics
- Sentence-level topic keywords (a_{st})
  - Obtain sentence-level topic distribution of each sentence in a_d
  - Merge topic keywords of the relevant topics for every sentence
- Retrieval
  - BM25F vs. Elasticsearch

BM25F

[Robertson & Zaragoza 2009]

$$\text{score}(q, a) = \sum_{t \in q \cap a} \left(\text{idf}(t) \times \frac{(k_3 + 1)c(t, q)}{k_3 + c(t, q)} \times \frac{(k_1 + 1)c'(t, a)}{k_1 + c'(t, a)} \right)$$

$$c'(t, a) = \frac{w_d \cdot c(t, a_d)}{1 - b + b\frac{|a_d|}{\bar{n}}} + w_{st} \cdot c(t, a_{st}) + w_{dt} \cdot c(t, a_{dt})$$

Elasticsearch

https://elastic.co/

– – –

$$\text{score}(q, a) = (1 - \text{tb}) \cdot \max_{f \in a} \text{score}(q, f) + \text{tb} \cdot \sum_{f \in a} \text{score}(q, f)$$

$$\text{score}(q, f) = \frac{1}{\sqrt{\sum_{t \in q} \text{idf}(t)^2}} \times \frac{|q \cap f|}{|q|} \times \sum_{t \in q} \left(\frac{\text{tf}(t, f) \cdot \text{idf}(t)^2 \cdot w_f}{\sqrt{|f|}} \right)$$

Experimental Setup

Dataset

[Park et al. 2015]

- 43,041 mobile applications
  - Extracted from Google Play
    https://play.google.com/
  - Name, description, category, reviews, etc.

- 56 non-exact queries

- Relevance of query-application pairs
  - Average of 81 judged applications per query
  - 3 judges per pair
  - No (0), partial (1), or perfect (2) satisfaction

Example: *“membership wallet”*

 Walmoo Wallet (2)

 LifeLock Wallet (0)

 MEO Wallet (0.667)

Evaluation

- Normalized Discounted Cumulative Gain (NDCG)
  [Järvelin & Kekäläinen 2002]
  - k = {3, 5, 10, 20}
  - Parameters tuned to maximize average NDCG@k

$$\mathrm{NDCG}_k = \frac{\mathrm{DCG}_k}{\mathrm{IDCG}_k} \qquad \mathrm{DCG}_k = \sum_{i=1}^{k} \frac{\mathrm{rel}_i}{\log_2(i+1)}$$

Results

		NDCG@3	NDCG@5	NDCG@10	NDCG@20
Description	BM25	0.569	0.540	0.523	0.537
	Elasticsearch	0.540	0.523	0.502	0.512
Topic	BM25F (w_{st}= 1, w_{dt}= 0)	0.554	**0.553**	**0.535**	0.530
	Elasticsearch (w_{st}= 2w_{dt}, tb = 0.1)	0.341	0.342	0.356	0.370
Description + Topic	BM25F (w_d= 0.96, w_{st}= 0, w_{dt}= 0.04)	**0.574**	0.542	0.527	**0.544**
	Elasticsearch (w_{st}= 2w_{dt}, w_d= w_{dt}, tb = 0.5)	0.552	0.532	0.504	0.519
References	LBDM [Wei & Croft 2006]	0.584	0.563	0.543	0.565
	Google Play https://play.google.com/	*0.589*	*0.575*	*0.568*	*0.566*

Discussion

Conclusions:

- Topics are relevant
  - Description-level
  - Sentence-level
- Topic keywords
  - Summarize description
  - Miss less common words

Future Work:

- Contextualized synonyms
- Review data
- Popularity boosting

Questions?

eugenio.ribeiro@inesc-id.pt

References

Blei, D.M. et al. (2003)
Latent Dirichlet Allocation
Journal of Machine Learning Research 3, pages 993-1022

Järvelin, K. & Kekäläinen, J. (2002)
Cumulated Gain-Based Evaluation of IR Techniques
ACM Transactions on Information Systems 20(4), pages 422-446

Park, D.H. et al. (2015)
Leveraging User Reviews to Improve Accuracy for Mobile App Retrieval
In SIGIR, pages 533-452

Robertson, S. & Zaragoza, H. (2009)
The Probabilistic Relevance Framework: BM25 and Beyond
Foundations and Trends® in Information Retrieval 3(4), pages 333-389

References

Wei, X. & Croft, W.B. (2006)
LDA-Based Document Models for Ad-Hoc Retrieval
In SIGIR, pages 178-185

Yi, X. & Allan, J. (2009)
A Comparative Study of Utilizing Topic Models for Information Retrieval
In ECIR, pages 29-41

Zhuo, J. et al. (2015)
Semantic Matching in APP Search
In WSDM, pages 209-210