TraininG towards a society of data-saVvy inforMation prOfessionals to enable open leadership INnovation

MOVING

Performance Comparison of Ad-hoc Retrieval Models over Full-text vs. Titles of Documents

Ahmed Saleh, Tilman Beck, Lukas Galke, Ansgar Scherp

ICADL 2018, Hamilton, New Zealand, 21 November 2018

www.moving-project.eu

Motivations

- **Question:** Can titles be sufficient for information retrieval task?

Query

Document collection

IR model

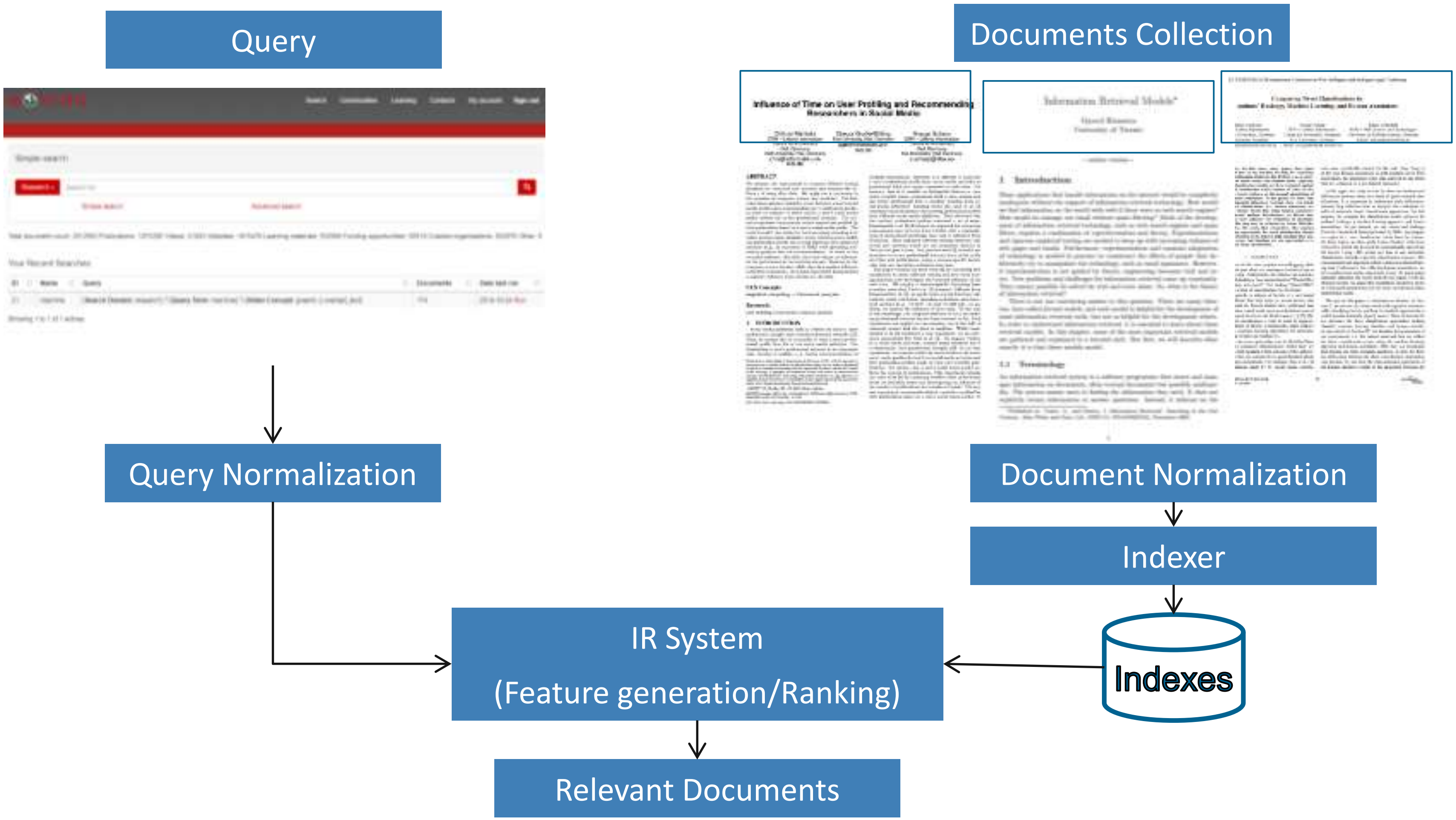
Relevant documents

Previous Studies [1]

MOVING
www.moving-project.eu

Authors	Title [Year]	Contribution:
Barker, Frances H and Veal, Douglas C and Wyatt, Barry K	Comparative Efficiency Of Searching Titles, Abstracts, and Index Terms In a Free-Text Database [1972].	Showed that Keywords can be searched more quickly than title material. The addition of keywords to titles increases search time by 12%, while the addition of digests increases it by 20%.
Lin, Jimmy	Is searching full text more effective than searching abstracts? [2009]	Lin used the MEDLINE test collection and two ranking models: BM25 and a modified TF-IDF in order to compare titles' retrieval vs. abstracts' retrieval.
Hemminger, Bradley M and Saelim, Billy and Sullivan, Patrick F and Vision, Todd J	Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts [2007]	- Comparing full-text searching to metadata (titles + abstract). - The authors used only an exact matching retrieval model to search for a small number of gene names in their study.

Overview


Query
Documents Collection
Influence of Time on User Profiling and Recommending Researchers in Social Media
Query Normalization
Document Normalization
Indexer
IR System
(Feature generation/Ranking)
Indexes
Relevant Documents

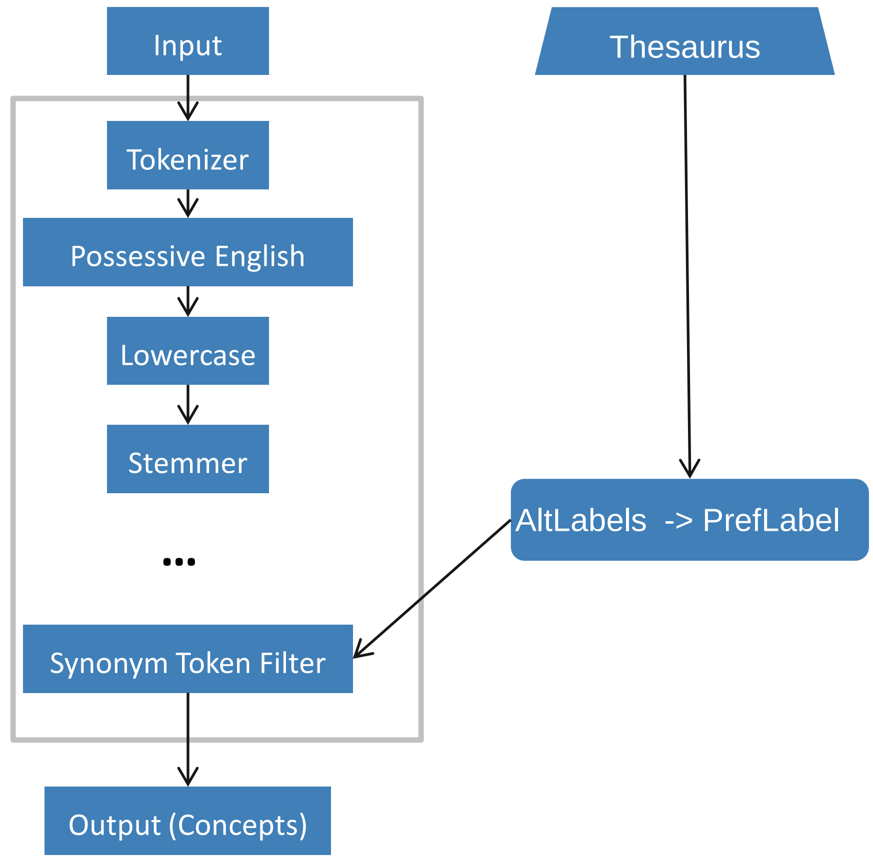
Query Normalization

- Preparing the query for semantics/statistic IR model.

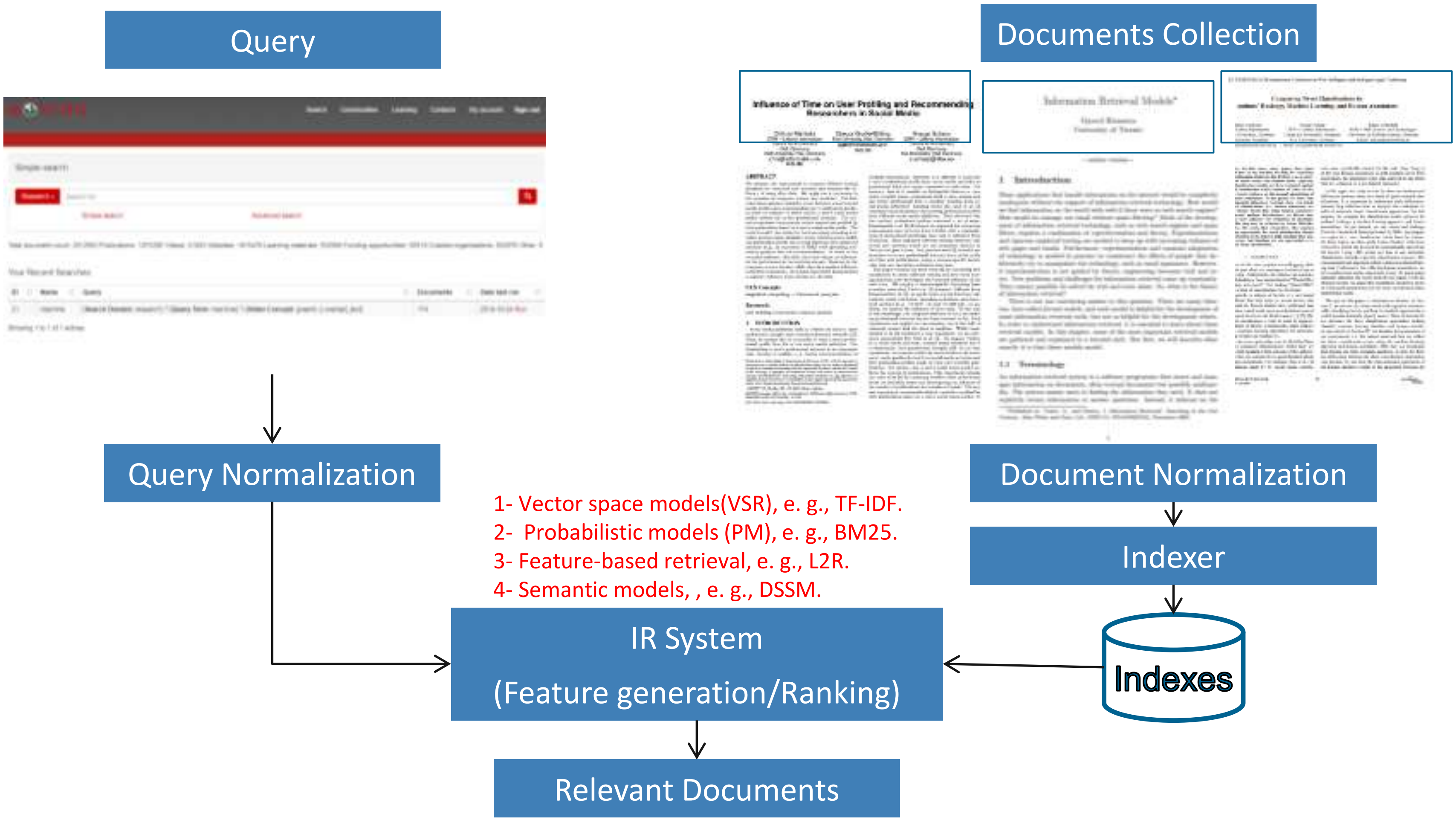
Query

Query Normalization

Example



Overall (recap)



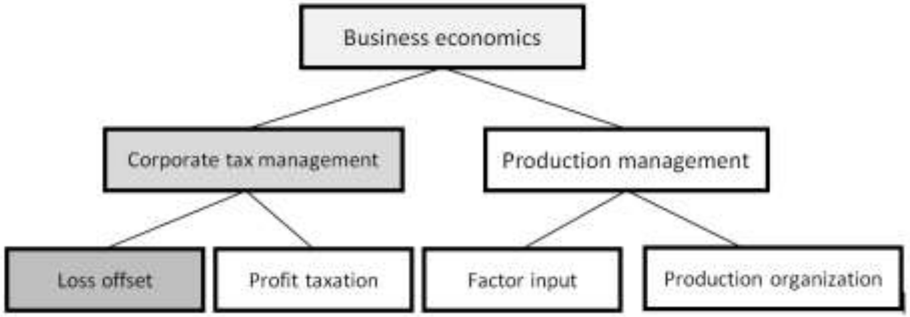
Compared models

- According to Croft et. Al [1], there are four main categories of ranking models:
  - Set theoretic models or Boolean models.
  - Vector space models(VSR), e. g., TF-IDF.
  - Probabilistic models (PM), e. g., BM25.
  - Feature-based retrieval, e. g., L2R.

- Furthermore, there are recent advances in Deep Learning that provide neural network IR models capable of capturing the semantics of words.
  - E.g. DSSM (Deep Structured Semantic Models) [2].

PM & VSR Models

- **Term Frequency – Inverse Documents Frequency (TF-IDF)**:
  - TF (w, d): is the number of occurrences of word w in documents d.
  - IDF: words that occur in a lot of documents are discounted (assuming they carry less discriminative information).
- **Okapi BM25:**
  - Another retrieval model which utilizes the IDF weighting for ranking the documents.
- **CF-IDF** is TF-IDF extension that counts concepts (e.g. STW) instead of terms
  - STW is the economics thesaurus provides a vocabulary of more than 6,000 economics' subjects
  - Developed and maintained by an editorial board of domain experts at ZBW
- **HCF-IDF** (Hierarchical CF-IDF)
  - Extract concepts which are not mentioned directly.



L2R models

MOVING
www.moving-project.eu

- Learning to Rank (L2R) is a family of machine learning techniques that aim at optimizing a loss function regarding a ranking of items.
  - L2R Features represents the relation between doc and query
  - L2R Features are Mostly are numbers (formulas, frequencies, …)

  For Example:

0	**qid:1**	1:0.000000	2:0.000000	3:0.000000	4:0.000000	5:0.000000 #**docid=30**
1	**qid:1**	1:0.031310	2:0.666667	3:4.00000	4:0.166667	5:0.033206 #**docid=20**
1	**qid:1**	1:0.078682	2:0.166667	3:7.00000	4:0.333333	5:0.080022 #**docid=15**

- L2R models fall into three categories:
  - **Pointwise models:** relevancy degree is generated for every single document regardless of the other documents in the results list of the query.
  - **Pairwise models:** considers only one pair of documents at a time (e.g. LambdaMart).
  - **Listwise models:** the input consists of the entire list of documents associated with a query (e.g. Coordinate Ascent)

Semantic Models (SM)

- **Deep Semantic Similarity model (DSSM)[4]:**
  - The model uses a multilayer feed-forward neural network to map both the query and the title of a webpage to a common low-dimensional vector space.
  - The similarity between the query-document pairs is computed using cosine similarity.

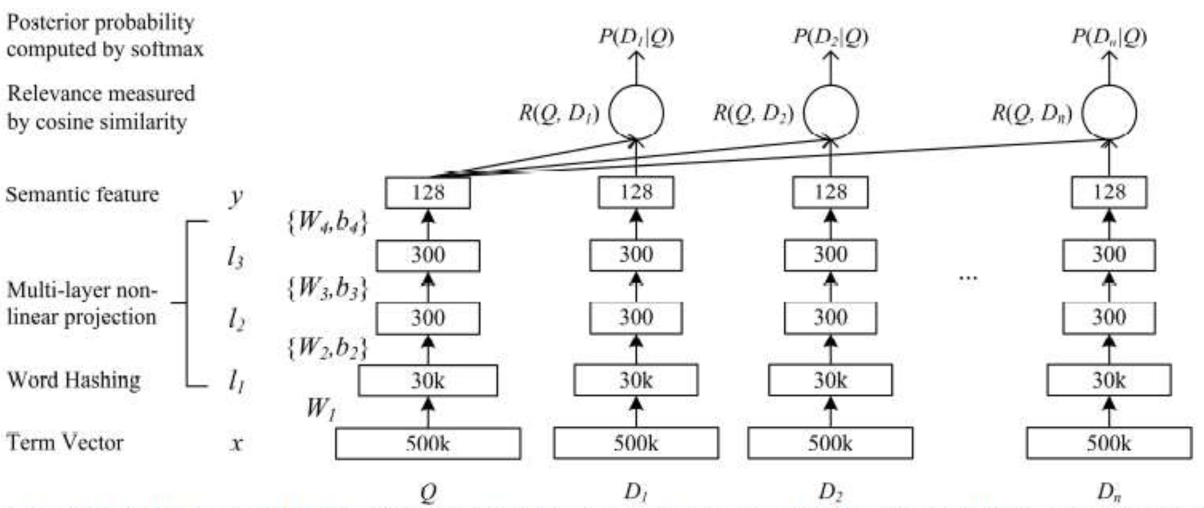
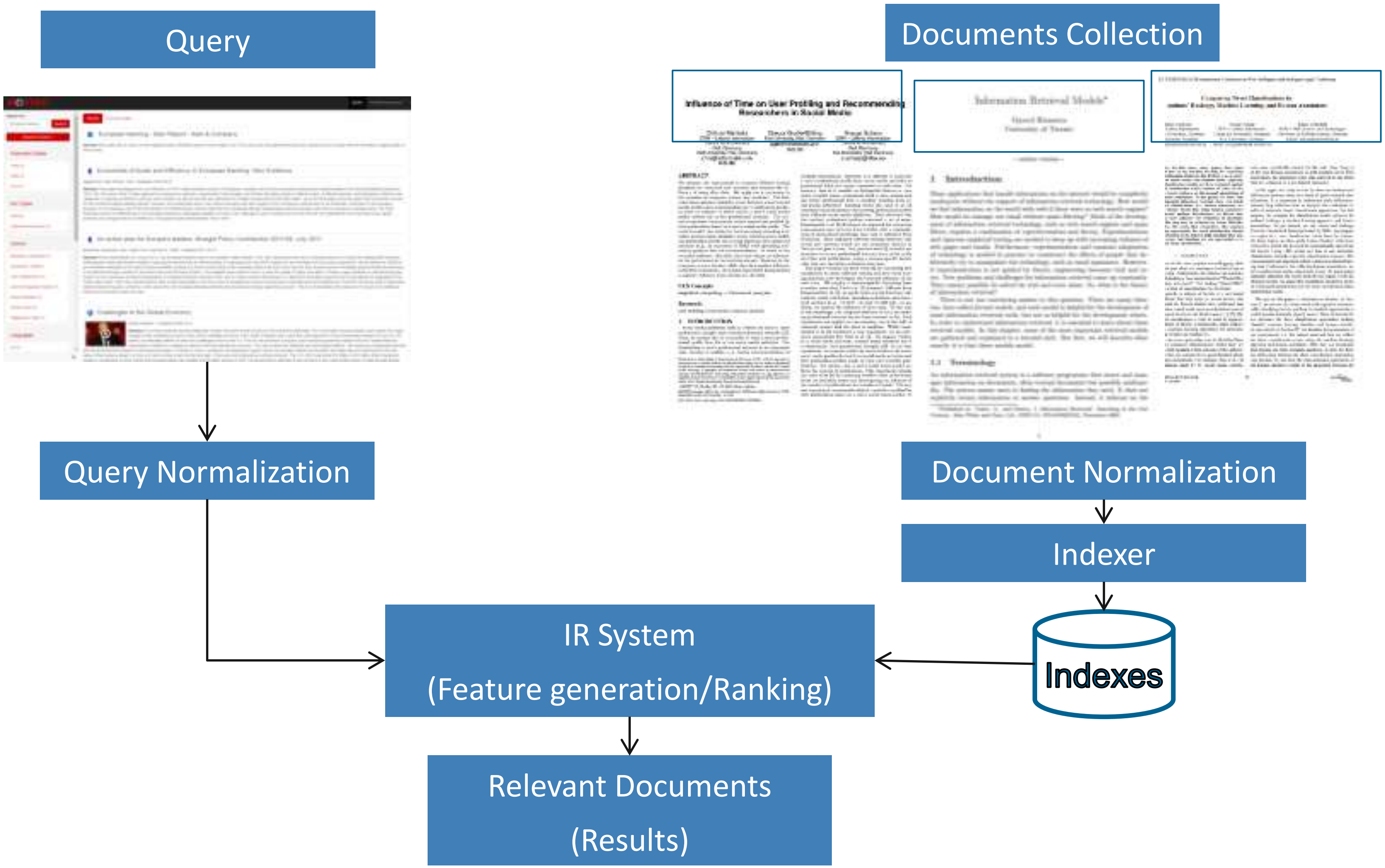


Figure 1: Illustration of the DSSM. It uses a DNN to map high-dimensional sparse text features into low-dimensional dense features in a semantic space. The first hidden layer, with 30k units, accomplishes word hashing. The word-hashed features are then projected through multiple layers of non-linear projections. The final layer's neural activities in this DNN form the feature in the semantic space.

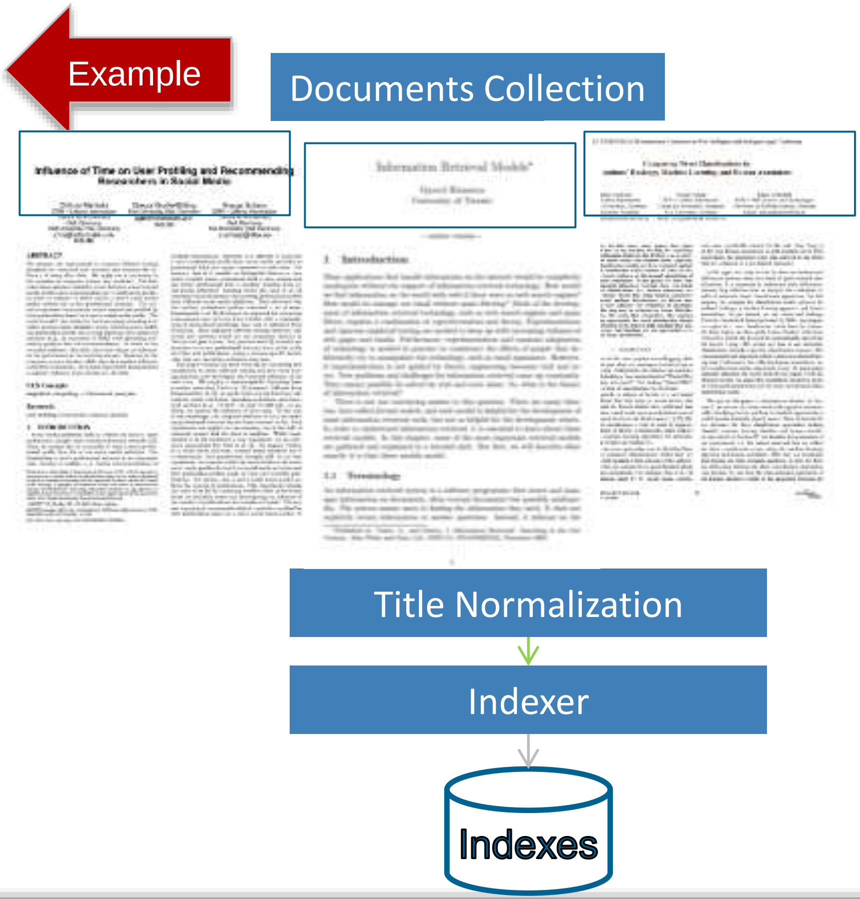
- **Convolutional Deep Semantic Similarity (C-DSSM)[5]**

Overall (recap)



Datasets (1)

- **The datasets are composed to two types: Labeled and Unlabeled.**
  - **Labeled datasets:** a document is given a binary classification as either relevant or non-relevant.
  - **Unlabeled datasets:** a hierarchical domain-specific thesaurus that provides topics (or concepts) of the libraries' domain is included.
  we consider the document as relevant to a concept if and only if it is annotated with the corresponding concept.

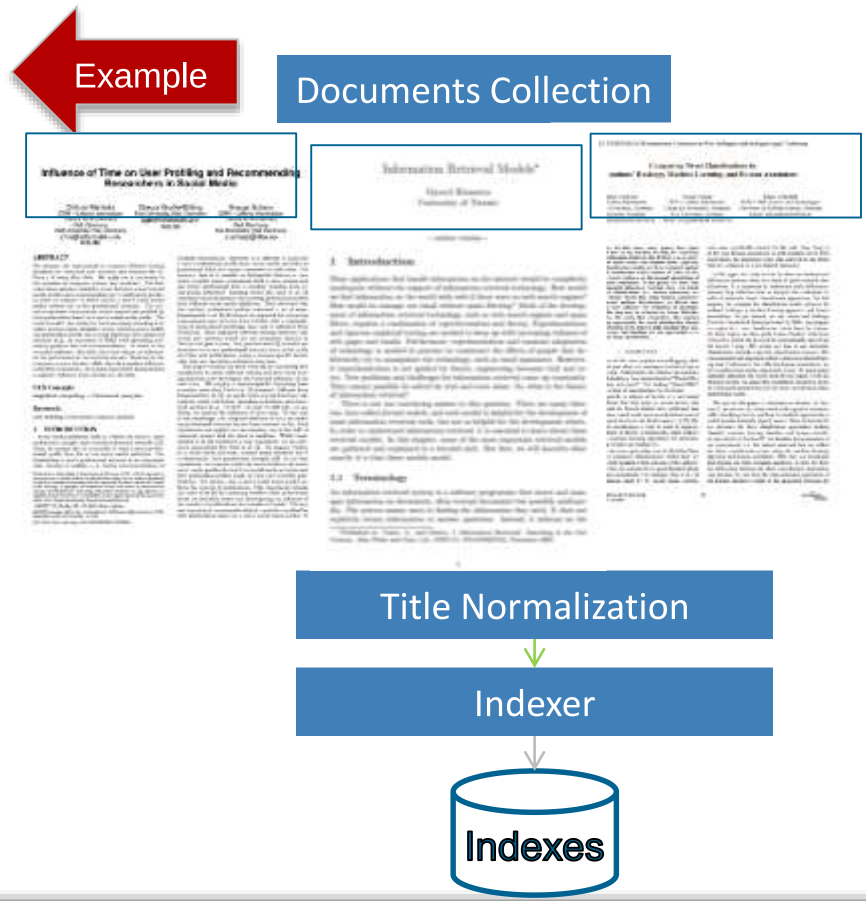


Datasets (2)

- The datasets are composed to two types: Labeled and Unlabeled.
- We used the following datasets:

		# of documents	# of queries	More information
Labeled Datasets	NTCIR-2[1]	322,059	49	consists of rel. Judgments of 66,729 pairs
	TREC[2]	507,011	50	consists of rel. Judgments of 72,270 pairs
Unlabeled Datasets	EconBiz[3]	288,344	6,204	Economics‘ scientific publications
	IREON[4]	27,575	7,912	Politics‘ scientific publications
	PubMed[5]	646,655	28,470	Bio-medical‘ scientific publications

1 http://research.nii.ac.jp/ntcir/permission/perm-en.html#ntcir-2
2 https://trec.nist.gov/data/intro_eng.html
3 https://www.econbiz.de/
4 https://www.ireon-portal.de/
5 https://www.ncbi.nlm.nih.gov/pubmed/



Comparison Results - labeled datasets

- With manual annotations as gold-standard.
- **Dataset:**

		# of documents	# of queries
Labeled Datasets	NTCIR-2	322,059	66,729
	TREC	507,011	72,270

- **Queries:**
  - short queries from the same dataset.
- **29 features for L2R:**
  - MK + Modified LETOR + Word2Vec + Ranking models.

- The metric $nDCG$ compares the top documents (DCG), with the gold standard and is computed as follows:
  - $nDCG_k = \frac{DCG_k}{IDCG_k}$ where $DCG_k = \mathrm{rel}_1 + \sum_{i=2}^{k} \frac{\mathrm{rel}_i}{Log(i)}$
  - D is a set of documents, $rel(d)$ is a function that returns one if the document is rated relevant, otherwise zero, and $IDCG\_k$ is the optimal ranking.

Comparison Results - labeled datasets

MOVING
www.moving-project.eu

Family	Method	NTCIR-2		TREC	
		Titles	Full-text	Titles	Full-text
VSM	TF-IDF	0.19	0.18	0.21	0.39
	CF-IDF	0.05	0.05	0.12	0.13
	HCF-IDF	0.23	0.24	0.10	0.12
PM	BM25	0.24	0.32	**0.23**	**0.41**
	BM25CT	0.24	0.31	0.20	0.405
L2R - FFS	L2R – LambdaMART	0.25	0.30	0.22	0.39
	L2R – RankNet	0.28	0.29	0.13	0.10
	L2R – RankBoost	0.26	0.32	0.21	0.34
	L2R – AdaRank	0.21	0.31	0.19	0.22
	L2R – ListNet	0.21	0.24	0.15	0.07
	L2R – Coord. Ascent	0.29	0.33	0.22	0.39
SM	DSSM	**0.33**	0.26	0.18	0.23
	C-DSSM	0.32	0.32	0.18	0.20
L2R – BFS	L2R – LambdaMART	0.20	0.15	0.16	0.33
	L2R – RankNet	0.28	0.15	0.05	0.046
	L2R – RankBoost	0.26	0.25	0.13	0.38
	L2R – AdaRank	0.29	**0.37**	0.18	0.37
	L2R – ListNet	0.29	**0.37**	0.29	0.37
	L2R – Coord. Ascent	0.29	**0.37**	0.29	0.38

Comparison Results - unlabeled datasets

- Dataset:

		# of documents	# of queries
Unlabeled Datasets	EconBiz	288,344	6,204
	IREON	27,575	7,912
	PubMed	646,655	28,470

- Gold-standard: Domain experts annotations.

- Queries:
  - ZBW's economics thesaurus.
  - FIV politics thesaurus.
  - MeSH labels, medical thesaurus.

Titles vs full text on unlabeled datasets

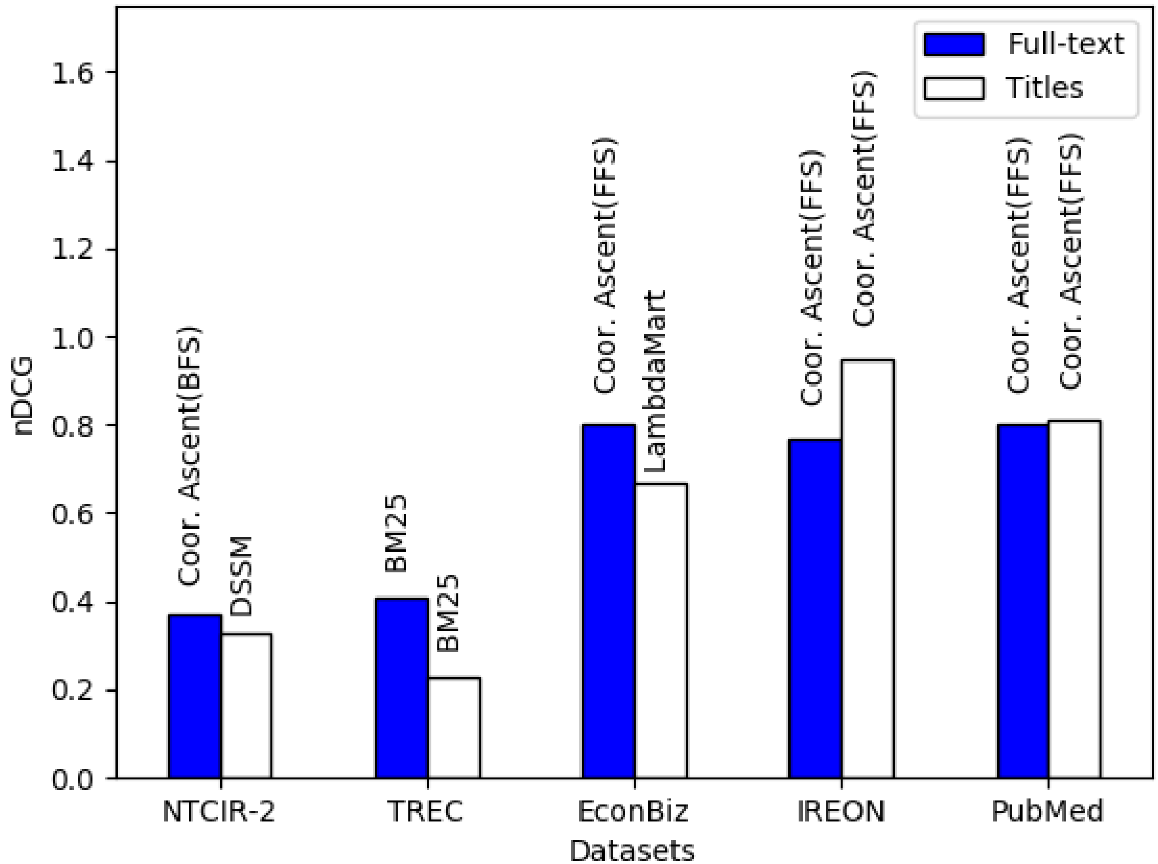
MOVING
www.moving-project.eu

Family	Method	EconBiz		IREON		PubMed	
		Titles	Full-text	Titles	Full-text		
VSM	TF-IDF	0.26	0.22	0.661	0.36	0.80	0.54
	CF-IDF	0.13	0.19	0.44	0.32	0.66	0.49
	HCF-IDF	0.25	0.20	0.659	0.37	0.80	0.54
PM	BM25	0.25	0.20	0.662	0.37	0.80	0.55
	BM25CT	0.27	0.19	0.660	0.37	**0.81**	0.56
L2R - FFS	L2R – LambdaMART	**0.67**	0.68	0.83	0.69	0.67	0.67
	L2R – RankNet	0.28	0.10	0.20	0.21	0.30	0.30
	L2R – RankBoost	0.52	0.69	0.80	0.59	0.70	0.79
	L2R – AdaRank	0.50	0.67	0.79	0.65	0.56	0.52
	L2R – ListNet	0.28	0.10	0.20	0.20	0.30	0.30
	L2R – Coord. Ascent	0.57	**0.80**	**0.95**	**0.77**	**0.81**	**0.80**
SM	DSSM	0.29	0.33	0.41	0.39	0.34	0.33
	C-DSSM	0.29	0.34	0.42	0.44	0.32	0.35
L2R – BFS	L2R – LambdaMART	0.56	0.63	0.70	0.65	0.42	0.65
	L2R – RankNet	0.28	0.10	0.26	0.41	0.59	0.63
	L2R – RankBoost	0.52	0.10	0.80	0.47	0.30	0.72
	L2R – AdaRank	0.48	0.49	0.94	0.41	0.59	0.79
	L2R – ListNet	0.28	0.28	0.94	0.41	0.39	0.49
	L2R – Coord. Ascent	0.53	0.10	0.94	0.69	0.59	0.78

Titles vs full text –results

MOVING
www.moving-project.eu

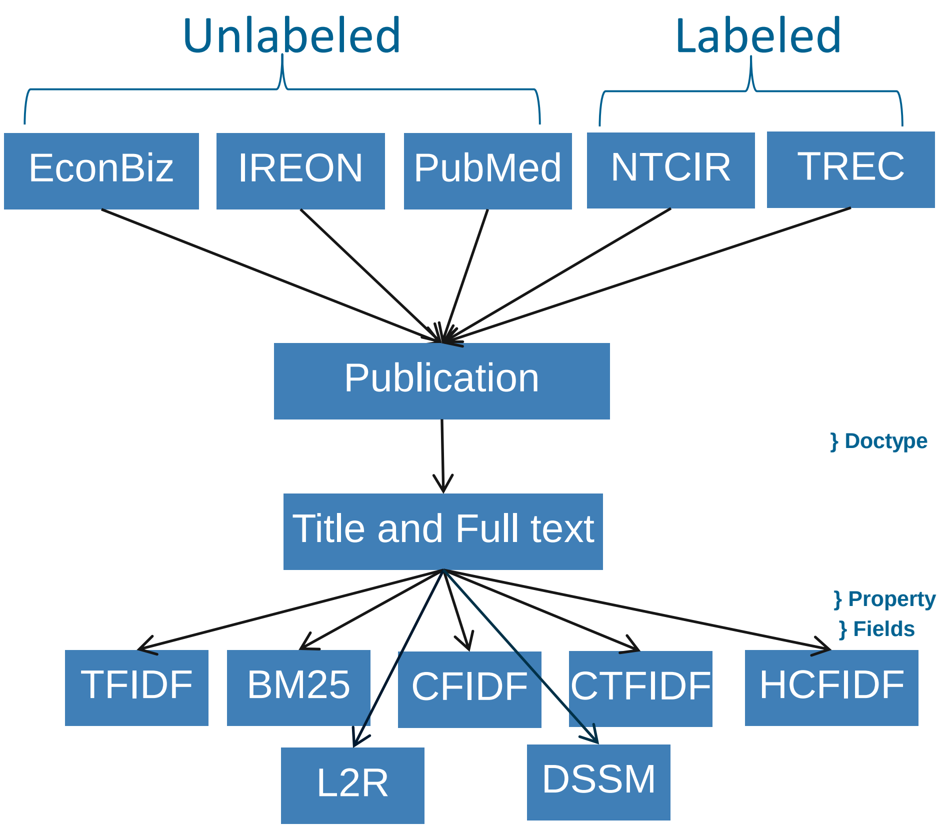
- Aggregating the best $nDCG$ values overall datasets and configurations. The best full-text-based retrieval models attains only 3% more than The best titles-based retrieval models.

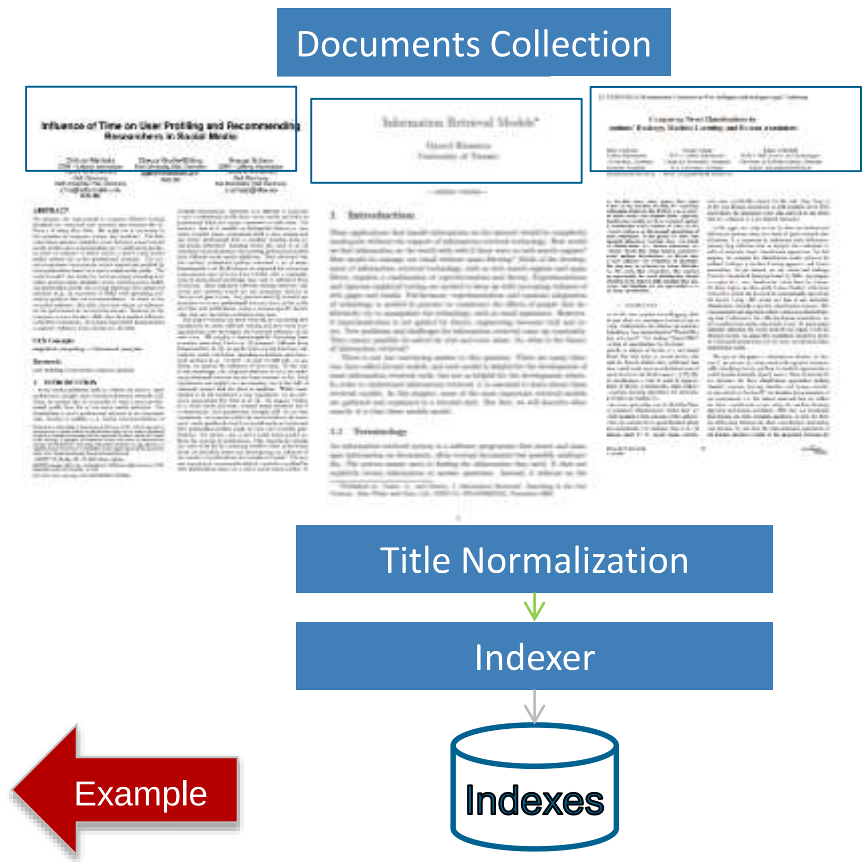


Replicate experiment results

MOVING
www.moving-project.eu

- Source code is available[1] .

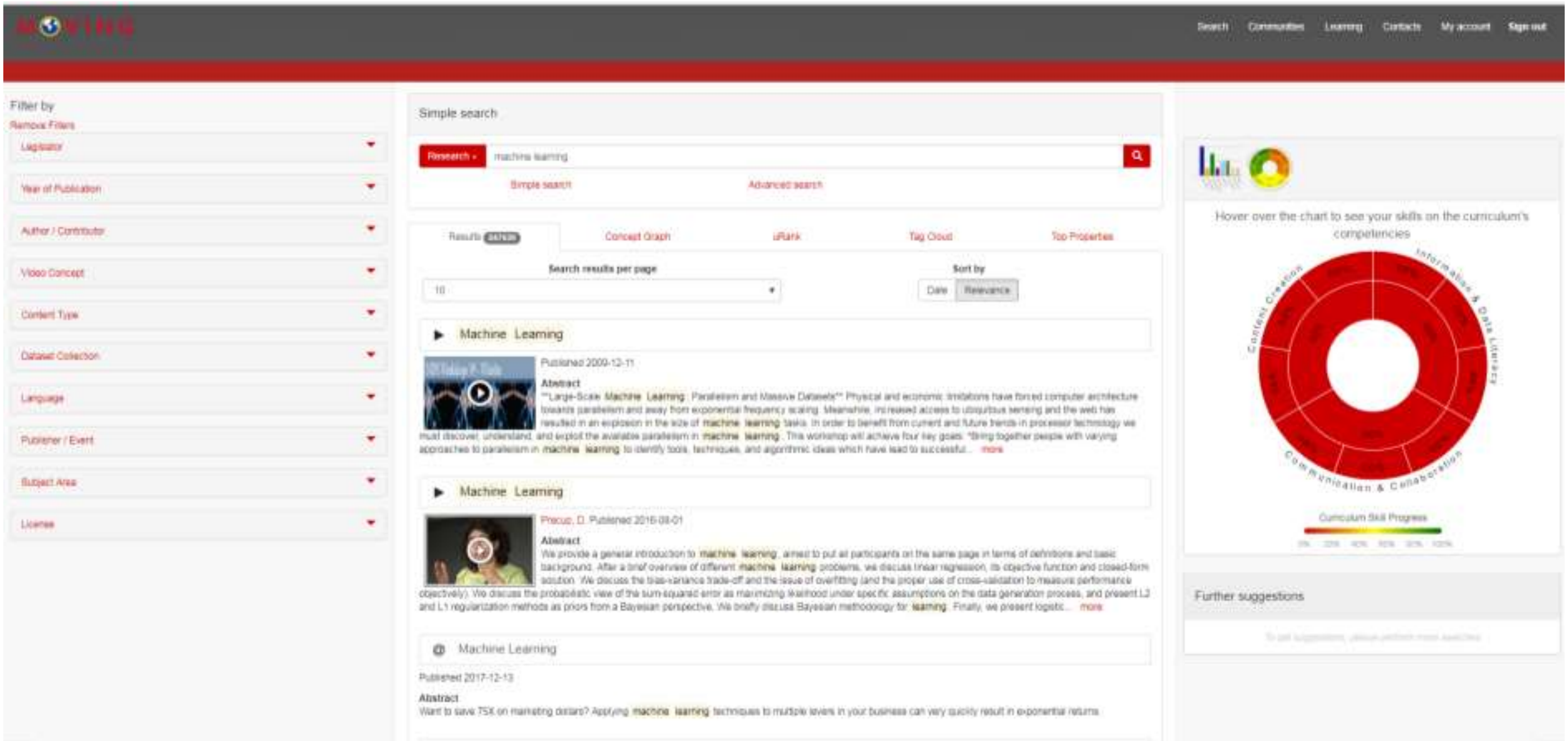




[1] https://bitbucket.org/a_saleh/icadl2018/src

MOVING Platform

- URL: http://platform.moving-project.eu



Conclusions:

- We conducted a study to compare title-based with full-text-based ad-hoc retrieval.

- We compared different retrieval models of different families (probabilistic models, vector space, learning to rank models and semantic models).

- We used five datasets, out of which three datasets are obtained from digital libraries: Econbiz, PubMed and IREON, and two standard test collections

- Our experiments show that title-based ad-hoc retrieval models can provide close, and sometimes even better, results compared to the full-text ad-hoc retrieval.

Project consortium and funding agency

MOVING is funded by the EU Horizon 2020 Programme under the project number INSO-4-2015: 693092

Thank you for your attention!

Any questions?

References:

MOVING
www.moving-project.eu

1. Croft, W. Bruce, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*. Vol. 283. Reading: Addison-Wesley, 2010.
2. Huang, Po-Sen, et al. "Learning deep structured semantic models for web search using clickthrough data." Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.
3. Huang, Po-Sen, et al. "Learning deep structured semantic models for web search using clickthrough data." Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.
4. Shen, Yelong, et al. "Learning semantic representations using convolutional neural networks for web search." Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014.

L2R models

- **Main L2R models:**
  - LambdaMart (Pairwise):
    - Combines LambdaRank, a neural network pairwise L2R approach, and Multiple Additive Regression Trees (MART), which uses gradient boosted decision trees for prediction.
    - When comparing a pair of documents, the gradient of the cost function indicates in which direction a document should move in a ranked list.
  - Coordinate Ascent (Listwise):
    - Optimization technique for unconstrained optimization problems
    - Scoring function is comprised of a linear combination of the features.
    - Optimizes the objective function by iteratively choosing one dimenstion (or feature) to search for, and fix all other dimensions

L2R features

- Represents the relation between doc and query
- Mostly are numbers (formulas, frequencies, …)
  - e.g. 0 **qid:1** 1:0.000000 2:0.000000 3:0.000000 4:0.000000 5:0.000000 #**docid=30**
    1 **qid:1** 1:0.031310 2:0.666667 3:4.00000 4:0.166667 5:0.033206 #**docid=20**
    1 **qid:1** 1:0.078682 2:0.166667 3:7.00000 4:0.333333 5:0.080022 #**docid=15**

Metzler and Kanungo - MK Set	Sentence length, Exact match, Term overlap, Synonym overlap, Language Model with Dirichlet smoothing
Modified LETOR	Covered query term number, IDF, Sum/Min/Max/Mean/Variance of TF, Sum/Min/Max/Mean/Variance of length normalized TF, Sum/Min/Max/Mean/Variance of TF-IDF, Language model absolute discounting smoothing, Language model with Bayesian smoothing using Dirichlet priors, Language model with Jelinek-mercer smoothing
Ranking model features	TF-IDF, BM25, CF-IDF, HCF-IDF, Word2Vec

L2R Best Feature Set (BFS)

MOVING
www.moving-project.eu

- A good IR system can retrieve the most important documents in a fast and scalable way using only a limited amount of information about the query and documents.

- **Goal:** find a meaningful subset of features which can still produce sound results.
  - Correlation-based Feature Selection algorithm (CFS)
  - The CFS algorithm computes a score for a subset S of the 29 features containing k features using the following equation

$$\text{score}_{\text{CFS(S)}} = \frac{k \cdot \overline{r_{gf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

    - Where r_{gf} is average gold standard g – feature f correlation
    - The formula denotes higher scores to the subsets which have a low 'feature-feature' correlations and high 'gold standard-feature' correlations.
    - We calculated $score\_CFS(S)$ for all feature subsets of sizes $|S| = \{1, \ldots, 29\}$, which equals $2^{29} - 1 = \mathbf{536{,}870{,}911}$ possible subsets.

L2R Best Feature Set (BFS)

MOVING
www.moving-project.eu

- The large table that includes the best featuresets.

Dataset	Content	Best Feature Set (BFS)	#	$Score_{CFS(s)}$
NTCIR-2	**Full-Text**	BM25, Exact match	2	0.20
	Titles	BM25, Exact match	2	0.15
TREC	**Full-Text**	BM25, Exact match, Sum of length normalized TF	3	0.28
	Titles	BM25, Language model with Dirichlet smoothing, Minimum of TF-IDF, Term overlap, Word2vec	5	0.13
EconBiz	**Full-Text**	Language model with absolute discounting smoothing, Language model with bayesian smoothing using Dirichlet priors, Min TF-IDF, Var TF-IDF	4	0.41
	Titles	BM25, Exact match, Language model, Synonym overlap, Term overlap, Covered query term number, Max TF-IDF, Mean length norm TF, Mean TF, Mean TF-IDF, Min length norm TF, Min TF, Min TF-IDF, Sum length norm TF, Sum TF, Sum TFIDF	16	0.71
Politics	**Full-Text**	Language model with Dirichlet smoothing, Language model with absolute discounting smoothing, Language model with Jelinek-Mercer smoothing, Max TF-IDF, Mean TF-IDF, Min TF-IDF, Sum TF, Sum TF-IDF, Var TF-IDF	9	0.41
	Titles	BM25	1	0.54
PubMed	**Full-Text**	Language model with Jelinek-Mercer smoothing, Mean TF-IDF	2	0.46
	Titles	Language model with absolute discounting smoothing, IDF	2	0.44