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

MOVING

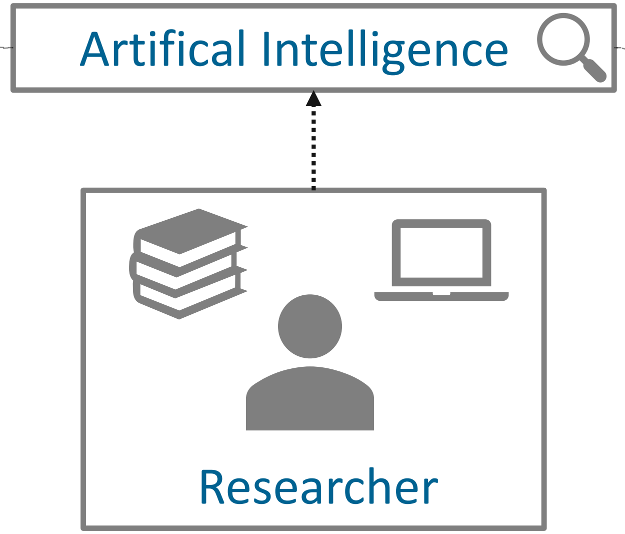
What to read next? Challenges and Preliminary Results in Selecting Representative Documents

TIR Workshop at DEXA 2018

Tilman Beck, Falk Böschen, Ansgar Scherp

www.moving-project.eu

Scenario: Broad Topic Search



evaluation of all results is **time-consuming**

classical ranking of result list is **inappropriate**

Machine Learning | Computer Vision | Robotics | …

users want to get an **overview of the content** (multiple subtopics)

document A = document A

users don‘t want to read **similar content** while getting an overview

Related Work and our Research Questions

Idea of selecting representative documents is not new:

- Zhang et al. [1] (2016) investigated how to find a **representative subset** from large-scale documents
  - representative subset: **high coverage** of original document set, **low redundancy** within subset, similar **content distribution** than superset
  - their approach: X-Means clustering + selection by coverage & redundancy
  - evaluation using a coverage and a redundancy measure

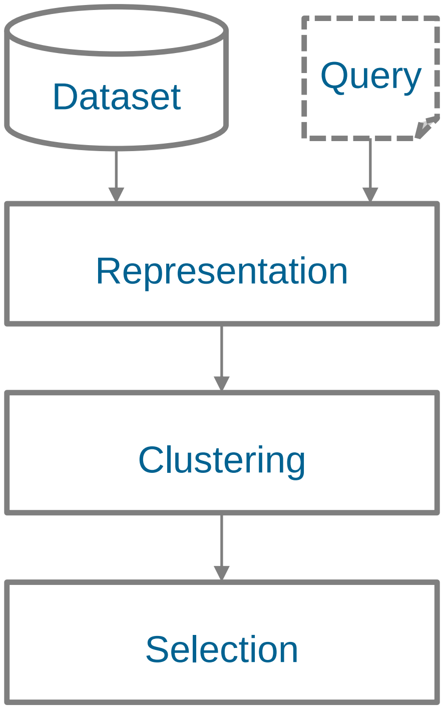
- We further investigated in this direction by extending their approach to answer the following research questions:
  - **RQ1**: What influence does the choice of a) document representation, b) clustering algorithm, and c) selection method have on the coverage and redundancy scores of the representative subset?
  - **RQ2**: Are the evaluation measures, coverage and redundancy, sufficient to evaluate the representativeness of a document set?

Our Approach: Document Selection

1. Retrieve relevant documents by sending the query to an IR system and compute suitable representations

2. Apply clustering to identify subtopics

3. Select the most representative documents from each cluster



Document Representations and Clustering

Comparing two different text document representations:

- Bag-of-Words (**BOW**)
- Paragraph Vectors (**D2V**) by Le and Mikolov (2014) [2]

.. and two different document clustering algorithms:

- Spherical K-Means (**KM**):
  - adaption of K-Means using cosine similarity as distance function
- Latent Dirichlet Allocation (**LDA**):
  - Probabilistic, generative model which identifies hidden topics in document corpus. We consider these topics as clusters
  - Input: term-document count matrix + number of topics ("clusters")
  - Output: document-topic matrix where each entry is a probability of a document belonging to a topic

Representative Document Selection

MOVING
www.moving-project.eu

Considering baseline + two selection methods:

- **Baseline**: random selection (**R**) of documents from each cluster
- Selection by coverage and redundancy (**CR**) is motivated by Zhang et al. [1]:
  - First, from each cluster, select document being closest to centroid (maximum coverage of cluster)
  - Subsequently, documents with lowest similarity to previously selected documents are selected (minimizing redundancy)
- Selection by User Intent (**IA**):
  - Introduced by Agrawal et al. (2009) [3] to increase diversity of topics among search results
  - Probability-based approach computing the relevance of documents to the query & the probability to satisfy any of the k topics
  - Originally used with LDA, but can be adapted to cluster setting

cluster proportion used to compute number of documents to be selected

Evaluation

- Two datasets of scientific publications:

Name	ACL Anthology Network	PubMed Open Access
full-text documents	22,486	646,513
queries	10 sampled from ACM CCS	10 sampled from MeSH
avg documents per query	1,500	1,100

- Evaluation measures:
  - Coverage: how much of dataset *D* is covered by a subset *S*:

$$\text{coverage}(S, D) = \frac{1}{|D|} \sum_{r \in D} (\max_{d \in S}(\text{sim}(d, r)))$$

  - Redundancy: redundant information in subset *S* is assessed by:

$$\text{redundancy}(S) = \sum_{d_i \in S} (\frac{1 - 1 \backslash \sum_{d_j \in S} sim(d_i, d_j)}{|S|})$$

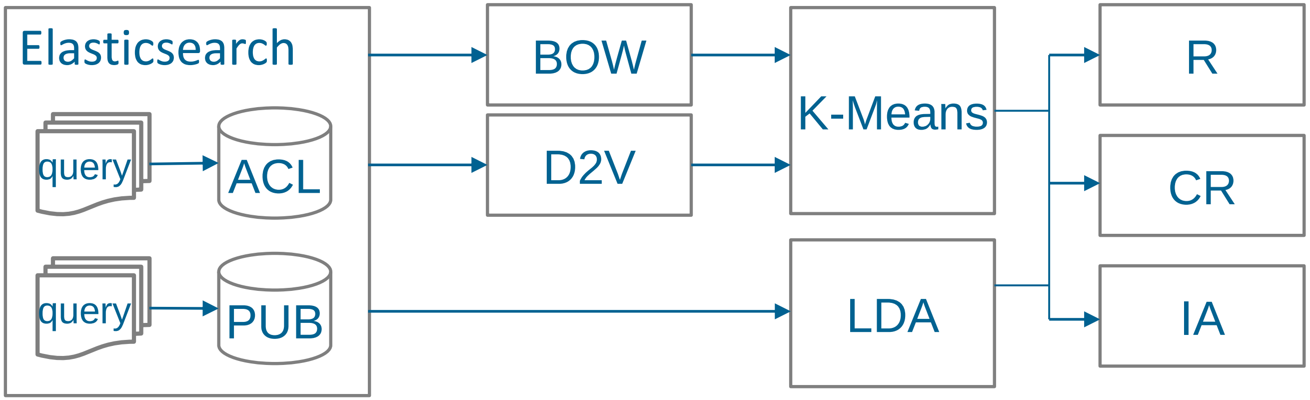
**sim* refers to cosine similarity between two documents

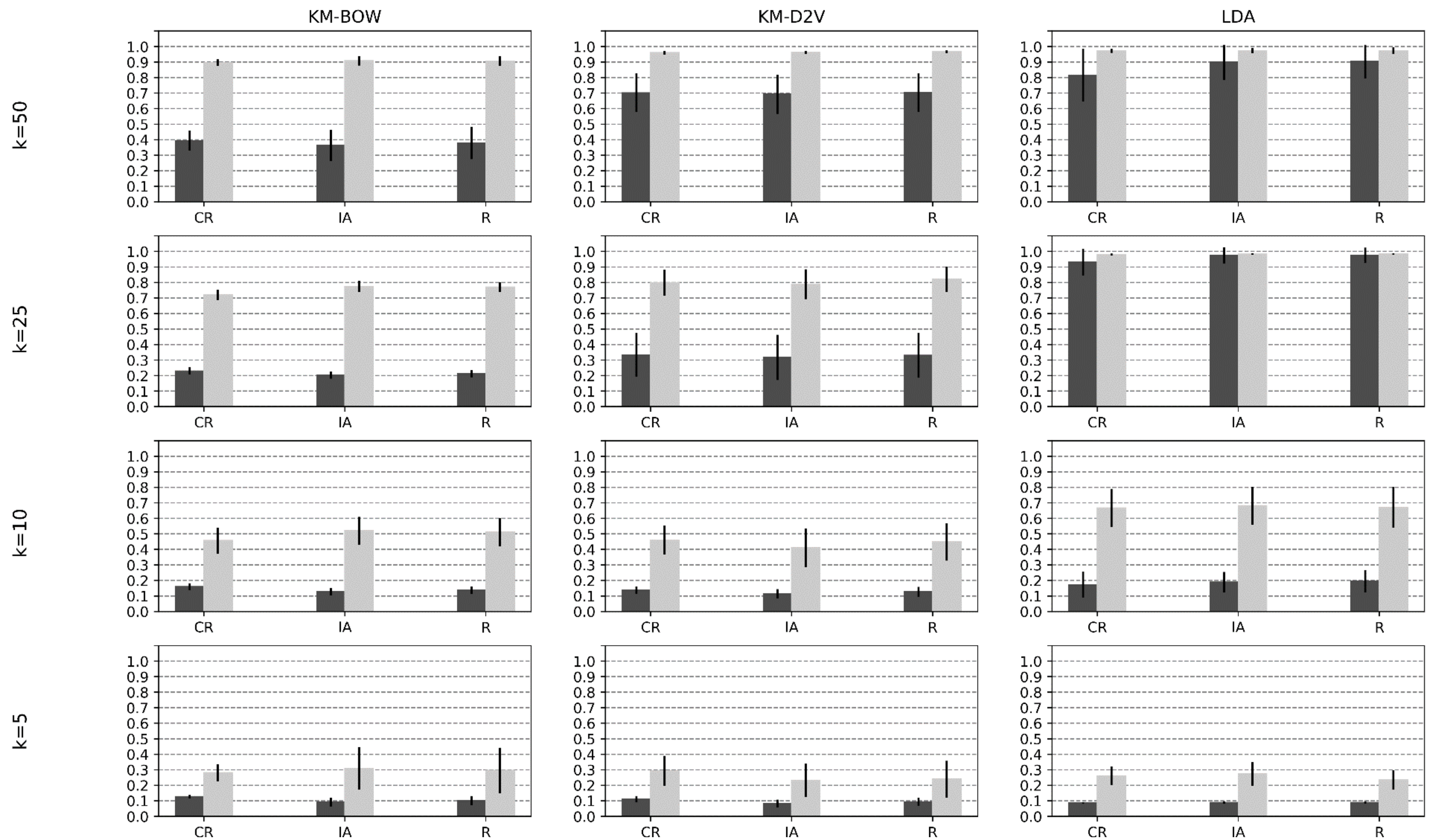
Experiment Setup

MOVING
www.moving-project.eu

- Procedure:
  - documents were prepocessed using Porter stemming, stop-word removal (NLTK), limitation of TF and vocab size
  - Representation computation:
    - BOW with BM25-weighting
    - D2V using model which was pre-trained on English Wikipedia dump
  - Clustering using different $k \in \{5,10,25,50\}$
  - Application of **CR**-Selection, **IA**-Selection and **R**andom-Selection

- In total 36 experiments:





Coverage (left black bars) and redundancy (right grey bars) averaged over all queries for the different document selection strategies on the ACL dataset using k ϵ {5,10,15,20}. The standard deviation is indicated as a black line on top of each bar.

Results for the first Research Question

MOVING
www.moving-project.eu

- **RQ1**: What influence does the choice of [...] have ?
  a) Document representation
    - for k=5 and k=10: no large difference for coverage, but for selection methods IA & R with document embeddings there is slightly less redundancy.
    - from k=25: selections based on KM-D2V have a higher coverage and a sharper increase in redundancy
    - Influence of D2V: for small *k* slightly less redundant content, for larger *k* more content is covered
  b) Clustering algorithm
    - except for LDA, coverage and redundancy results increase steadily, more distinct with larger *k*
    - LDA, from *k*=25, both measure scores close to 1
  c) Selection method
    - generally lower redundancy with CR + KM-BOW, less pronounced for larger *k*
    - poor performance of CR with KM-D2V and LDA with regards to redundancy
    - coverage: selection method less important than clustering algorithm

Results for the second Research Question

MOVING
www.moving-project.eu

RQ2: Are the evaluation measures, coverage and redundancy, sufficient to evaluate the representativeness of a document set?

- We made three interesting observations:

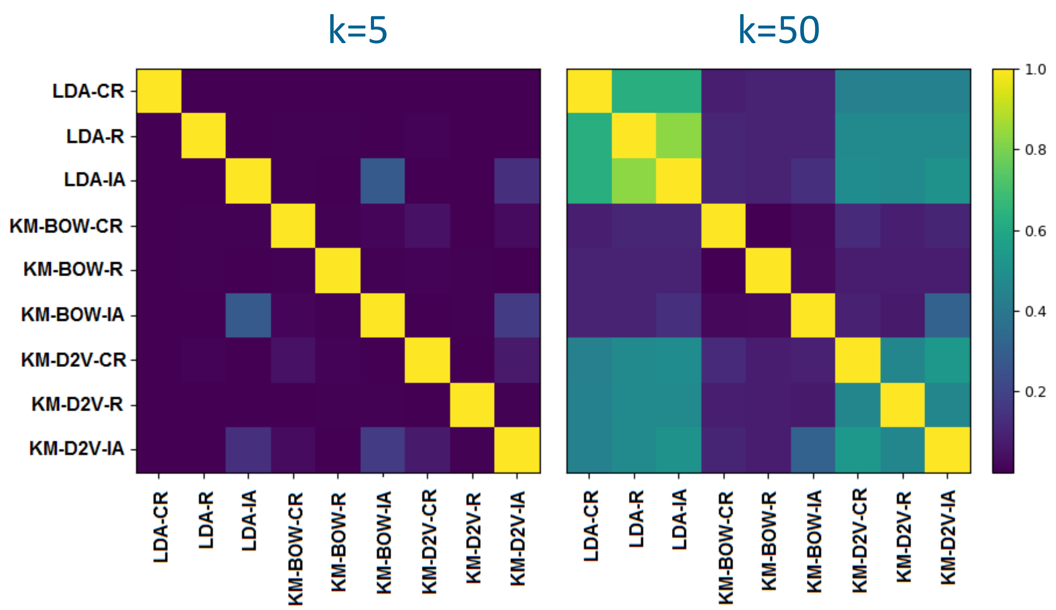
  1. Scores for both measures increase consistently for larger *k*
     - direct correlation with number of selected documents and, thus, with cluster proportion calculation
     - selection of more documents caused by heterogeneous clusters, which, in turn, are more likely for larger *k*
       → coverage and redundancy are inflated!

  2. For each strategy, redundancy exceeds coverage
     - in contrast to findings of Zhang et al. [1]
     - not caused by IR setting

Results for the second Research Question

M VING
www.moving-project.eu

3. Independence of evaluation measures from actual choice of documents
   - Random selection as baseline achieves comparable results as other strategies



   - Analysis of shared documents between strategies highlights that only with larger *k* more documents are in common
     → limits the generalization of coverage and redundancy to evaluate representativeness

Conclusion

- We proposed a **document selection framework** in an IR context
- There is **no unique representative document set** based on current evaluation measures → coverage and redundancy are insufficient
- Current computation of size of result set is error-prone (e.g. heterogeneous cluster sizes) and leads often to selection of too many documents

Resources:

1. Zhang, J., Liu, G., Ren, M.: Finding a representative subset from large-scale documents. J. Informetrics 10(3), pp. 762-775 (2016)
2. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014. JMLR Workshop and Conference Proceedings, vol. 32, pp. 1188-1196
3. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Baeza-Yates, R.A., Boldi, P., Ribeiro-Neto, B.A., Cambazoglu, B.B. (eds.) Proceedings of the Second International Conference onWeb Search andWeb Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009. pp. 5-14 ACM (2009)
4. Whissell, J.S., Clarke, C.L.A.: Improving document clustering using Okapi BM25 feature weighting. Inf. Retr. 14(5), pp. 466-487 (2011)
5. Endo, Y., Miyamoto, S.: Spherical k-means++ clustering. In: Torra, V., Narukawa, Y. (eds.) Modeling Decisions for Articial Intelligence - 12th International Conference, MDAI 2015, Skövde, Sweden, September 21-23, 2015, Proceedings. Lecture Notes in Computer Science, vol. 9321, pp. 103-114
6. Ma, B., Wei, Q., Chen, G.: A combined measure for representative information retrieval in enterprise information systems. J. Enterprise Inf. Management 24(4), pp. 310-321 (2011)

Project consortium and funding agency

iti Information Technologies Institute

EY Building a better working world

TECHNISCHE UNIVERSITÄT DRESDEN

KNOW Center

Institut "Jožef Stefan" Ljubljana, Slovenija

ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics

MANCHESTER 1824

gesis Leibniz-Institut für Sozialwissenschaften

PROGRESS & BUSINESS FOUNDATION

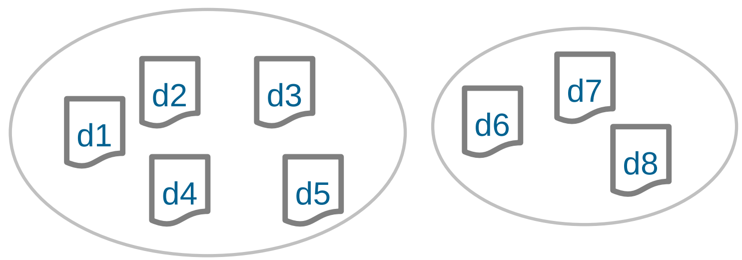
MOVING is funded by the EU Horizon 2020 Programme under the project number INSO-4-2015: 693092

Thank you for your attention!

Appendix A

- Number of documents to be selected r_i from each cluster c_i is dependent on proportion p_i of a cluster c_i :
  - $p_i = \frac{|ci|}{D}$
  - $r_i = \left\lfloor \frac{p_i}{p_{min}} \right\rfloor$

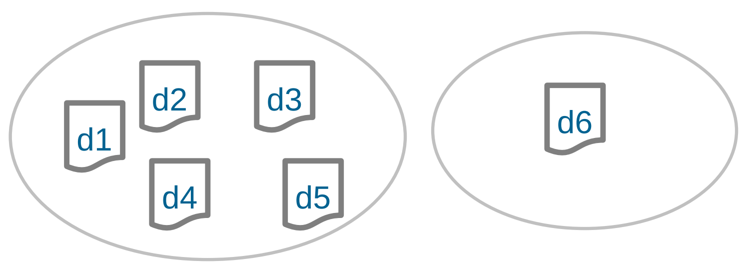
Example:



$$p_1 = \frac{5}{8}, \quad p_2 = \frac{3}{8}$$

$$r_1 = \left\lfloor \frac{5}{8} * \frac{8}{3} \right\rfloor = 2,$$

$$r_2 = \left\lfloor \frac{3}{8} * \frac{8}{3} \right\rfloor = 1$$



$$p_1 = \frac{5}{6}, \quad p_2 = \frac{1}{6}$$

$$r_1 = \left\lfloor \frac{5}{6} * \frac{6}{1} \right\rfloor = 5,$$

$$r_2 = \left\lfloor \frac{1}{6} * \frac{6}{1} \right\rfloor = 1$$

Appendix B

sim	d1	d2	d3	d4
d1	1.0	0.2	0.4	0.1
d2	0.2	1.0	0.8	0.0
d3	0.4	0.8	1.0	0.5
d4	0.1	0.0	0.5	1.0

- Coverage and redundancy measures:
  - $D = \{d1, d2, d3, d4\}$
  - $cov(\{d1, d2\}, D) =$

    $= \frac{1}{4}(1.0 + 1.0 + 0.8 + 0.1) = 0.725$
  - $red(\{d1, d2\}) = \frac{1}{2}\left(\left(1 - \frac{1}{1+0.2}\right) + \left(1 - \frac{1}{1+0.2}\right)\right) = 0.17$
- Edge cases:
  - $cov(\{d1, d2, d3, d4, d5\}, D) = \frac{1}{4}(1.0 + 1.0 + 1.0 + 1.0) = 1.0$
  - $red(\{d2, d4\}) = \frac{1}{2}\left(\left(1 - \frac{1}{1+0}\right) + \left(1 - \frac{1}{1+0}\right)\right) = 0$

Appendix C

- Dataset Statistics:

Statistic	ACL	PubMed
Storage Space	1.6gb	47.3gb
# of documents	22,486	646,513
$\|D_q\|$	1,524	1,101
$\|d\|$	2,638.01 (142.24)	2,166.89 (279.09)
$\|V_q\|$	31,356.20 (5,910.68)	18,640.50 (2,157.81)
Sparseness*	0.97	0.97

- $|D_q|$: size of the retrieved document set, averaged over all queries
- $|d|$: average document length (std deviation)
- $|V_q|$: average vocabulary size (after tuning)

* Sparseness was computed by dividing the number of zero entries in the document-term matrix by its size