RECLAK: ANALYSIS AND RECOMMENDATION OF INTERLINKING DATASETS

Giseli Rabello Lopes (PUC-Rio),
Bernardo Pereira Nunes (PUC-Rio),
Luiz André P. Paes Leme (UFF),
Marco A. Casanova (PUC-Rio)

Summary

- Motivation
- Background Information
- Proposed Technique
- Experiment
- Conclusions

Motivation

- Most of the data publishers typically link their datasets only to popular ones (DBpedia, Freebase, etc.)
  - difficulty of finding related datasets
  - strenuous task of discovering instance mappings
- We were interested in developing techniques that could find related datasets without instance mappings
- LAK as a test case
  - *For each dataset d_i, published in the LOD, is it interesting for the LAK administrator to try to link his dataset with d_i?*

Background Information

- We refer to the problem of finding datasets to link with as a recommendation problem as following

  - Given a finite set of datasets ***D*** and a dataset t, compute a rank score for each dataset $d_i \in \boldsymbol{D}$, denoted by $\boldsymbol{score(d_i,t)}$, such that the rank score of d_i increases with the chances of d_i being relevant for t.

  - Obs.: *to be relevant* means that one can find links between t and d_i

Background Information

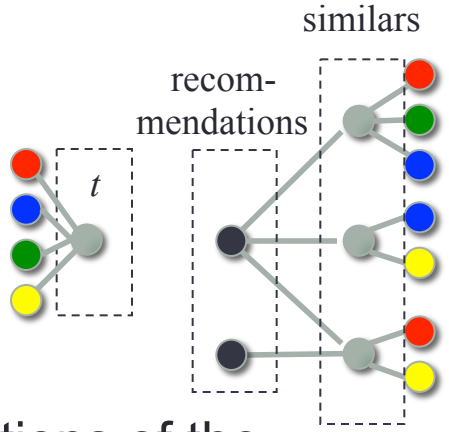
- *link*
  - An RDF *link* is an RDF triple whose subject and object are described in different datasets. (VoID Vocabulary)
- *linkset*
  - A *linkset* is a collection of such RDF links between two datasets. It is a set of RDF triples where all subjects are in one dataset and all objects are in another dataset. (VoID Vocabulary)
    - The subjects of all link triples are in a dataset referred as *subjectsTarget*
    - The objects of all link triples are in a dataset referred as *objectsTarget*
- If there is a *linkset* with $subjectsTarget = A$ and $objectsTarget = B$ then we say that:
  - *The dataset A is linked to the dataset B*
  - *The dataset B is linked from the dataset A*
  - *B is a connection of A*

Background Information

- *feature*
  - A *feature* of a dataset denotes a piece of data which is relevant to understand the content of the dataset
    - Class URIs
    - Property URIs
    - Vocabulary namespaces
    - objectsTarget URIs
    - User tags/topics
    - etc.
- *feature set*
  - A *feature set* of a dataset t denoted by $\boldsymbol{Bag}_t$ is a subset of the set of all possible features of t.

Recommendation Strategies

1. Semantic indexes (Nicolov et al. 2012)
2. Similarity
   - Recommend to a dataset t similar datasets
3. Global Popularity
   - Recommends to a dataset t the datasets most popular
4. Local popularity
   - Popular datasets in each information domain
5. "Friends of similar"
   - Recommends to a dataset t the datasets that are connections of the datasets similar to t.
     - Intuition by analogy with people's friendship: if there are people similar to me in terms of my personal interests then their friends can be relevant to me as well.

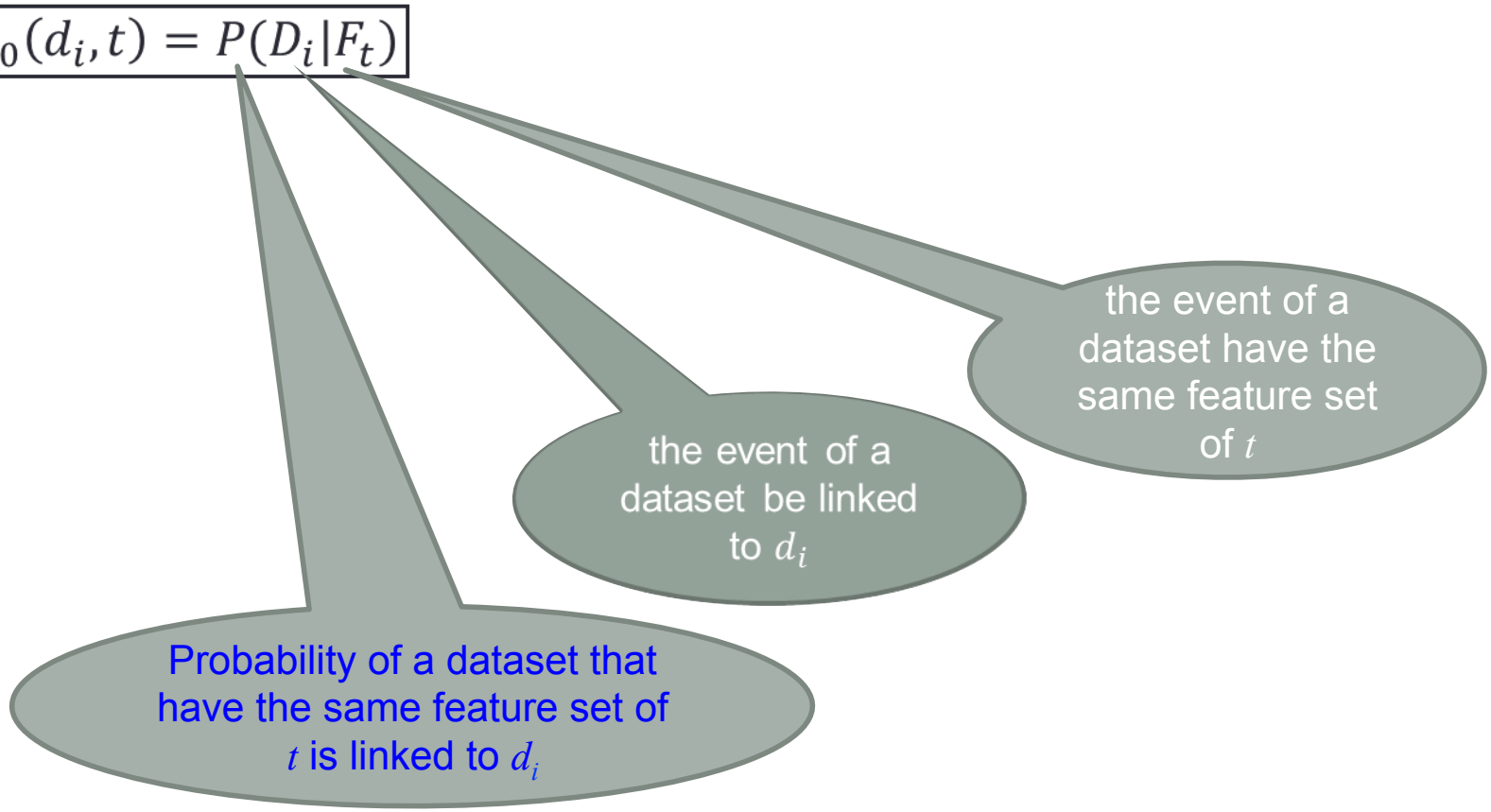


Recommendation techniques

- Bayesian ranking
  - $\boxed{score_0(d_i, t) = P(D_i|F_t)}$

Recommendation techniques

- Bayesian ranking
  - $score_0(d_i, t) = P(D_i|F_t)$



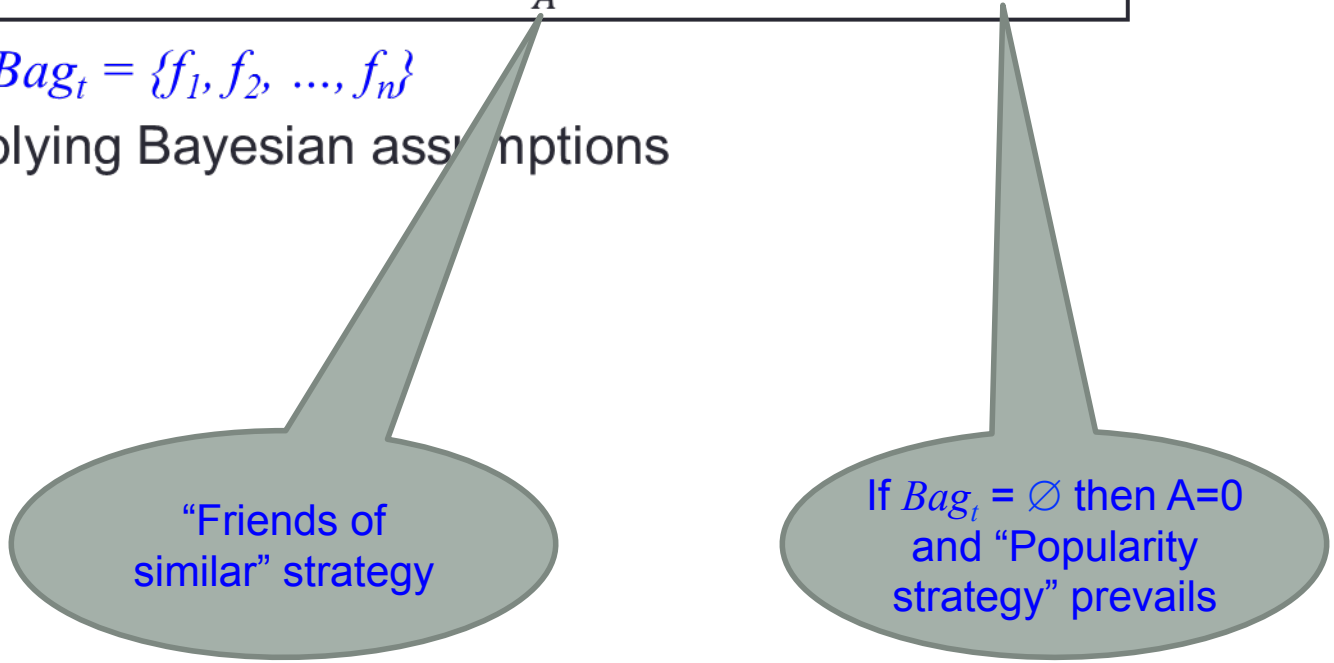
Recommendation techniques

- Bayesian ranking
  - $score(d_i, t) = \underbrace{\left(\sum_{f_j \, in \, Bag_t} logP\left(F_j \middle| D_i\right)\right)}_{A} + \underbrace{logP(D_i))}_{B}$
  - $t \cong Bag_t = \{f_1, f_2, \ldots, f_n\}$
  - Applying Bayesian assumptions

the event of a dataset have feature $f_j \in Bag_t$

Recommendation techniques

- Bayesian ranking
  - $score(d_i, t) = \underbrace{\left(\sum_{f_j \, in \, Bag_t} logP\left(F_j \middle| D_i\right)\right)}_{A} + \underbrace{logP(D_i))}_{B}$
  - $t \cong Bag_t = \{f_1, f_2, \ldots, f_n\}$
  - Applying Bayesian assumptions

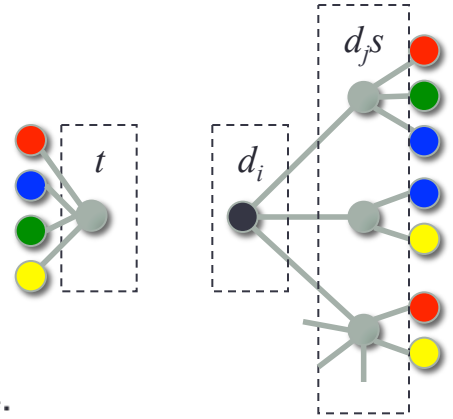


Recommendation techniques

- Bayesian ranking
  - Computation of probabilities
    - $P(F_j|D_i) = \frac{count(f_j,d_i)}{\sum_{j=1}^{m} count(f_j,d_i)} = \frac{\begin{matrix}number\ of\ datasets\ that\ \ have \\ feature\ f_j\ and\ objectsTarget\ d_i\end{matrix}}{total\ number\ of\ cooccurences\ (f_j,d_i)}$
    - $P(D_i) = \frac{count(d_i)}{\sum_{i=1}^{n} count(d_i)} = \frac{\begin{matrix}number\ of\ datasets\ that\ have\ d_i \\ as\ objectsTarget\end{matrix}}{total\ number\ of\ targets}$

Recommendation techniques

- Social Network-based ranking
  - $score(d_i, t) = ra(d_i, t) + \frac{pa(d_i,t)}{|D|}$
    - $t \cong Bag_t = \{f_1, f_2, \ldots, f_n\}$
    - $pa(d_i, t)$ = number of datasets linked to d_i
    - $|D|$ = number of available (known) datasets
    - $ra(d_i, t) = \sum_{d_j \in C} \frac{1}{popularity(d_j)}$
      - C = set of datasets that are similar to t and that are linked to d_i.
      - the relevance is inverse proportional to the popularity
      - Rationale:
        - if someone similar to me is very popular such as artists, politicians, etc. it is not guaranteed that their friends are also relevant to me.
        - if that person is not very popular, it may happen that the person is a friend or a colleague of mine and that their friends are relevant to me.



Recommendation techniques

- Social Network-based ranking
  - $score(d_i, t) = ra(d_i, t) + \frac{pa(d_i,t)}{|D|}$
    - $t \cong Bag_t = \{f_1, f_2, \ldots, f_n\}$
    - $pa(d_i, t)$ = number of datasets linked to d_i
    - $|D|$ = number of available (known) datasets
    - $ra(d_i, t) = \sum \frac{1}{popularity(d_j)}$
      - C = set of datasets that are similar to t and that are linked to d_i.
      - the relevance is proportional to the popularity
      - very popular such that it is not guaranteed that it is relevant to me.
        - if that person is not very popular, it may happen that the person is a friend or a colleague of mine and that their friends are relevant to me.

"Friends of similar" strategy

If $Bag_t = \varnothing$ then A=0 and "Popularity strategy" prevails

Recommendation techniques

- Performance
  - Previous work assessed the performance of the techniques using as Gold Standard the connections of datasets available in the Datahub.
  - Previous work demonstrated that, in average, the relevant datasets fits among the 20% datasets at the top of the ranking
    - We argue, therefore, that the techniques reduce 80% of the search space.

Experiment

Question

For each dataset d_i, published in the LOD, is it interesting for the LAK administrator to try to link his dataset with d_i?

- D = subset of datasets in Datahub (132 datasets)
- $features$ = set of class and property URIs used in all datasets in D (31 URIs)
- Performance
  - The experiment did not aimed at effectively finding links.
  - The performance of the techniques were assessed in previous work.
  - The rankings were assessed by inspection.

Experiment

Fragment of LAK features by number of datasets using the features

feature	# of datasets
Class	
foaf:Person	34
foaf:Organization	23
Property	
dce:title	61
dce:creator	57
foaf:name	42
foaf:homepage	37
dct:subject	24
dce:subject	12
foaf:member	9
foaf:based_near	8
foaf:lastName	7
foaf:mbox_sha1sum	7
foaf:maker	6
foaf:firstName	6

http://www.inf.puc-rio.br/~grlopes/RecLAK

Experiment

Datasets by used features

dataset	foaf:Organization	foaf:Person	swc:ConferenceEvent	swrc:InProceedings	swrc:Proceedings	bibo:authorList	dce:creator	dce:subject	dce:title	dct:subject	foaf:based_near	foaf:firstName	foaf:homepage	foaf:lastName	foaf:made	foaf:maker	foaf:mbox_sha1sum	foaf:member	foaf:name
acorn-sat							X		X										
aksworg	X	X							X	X			X					X	X
arrayexpress_e-mtab-104	X	X																	X
beneficiaries-of-the-european-commission										X			X						X
bibsonomy							X	X											
bizkaisense									X										
business_terms							X		X										
cablegate													X						
camera-deputati-linked-data									X										
chronicling-america								X											
courts-thesaurus										X									
datagov-catalog													X						
dbtune-john-peel-sessions		X							X										X
dcs-sheffield		X											X		X	X		X	X
debian-package-tracking-system																			X

http://www.inf.puc-rio.br/~grlopes/RecLAK

Experiment

#	Bayesian ranking	score*	#	SN-based ranking	score
1	semanticweb-org	-162.025	1	geonames-semantic-web	13.738
2	w3c-wordnet	-162.236	2	nytimes-linked-open-data	3.558
3	tags2con-delicious	-163.025	3	gnoss	3.051
4	dcs-sheffield	-163.025	4	lcsh	3.017
5	linked-open-camera	-163.025	5	rkb-explorer-acm	2.430
6	sweto-dblp	-163.025	6	rkb-explorer-wiki	2.408
7	geonames-semantic-web	-3281.339	7	dnb-gemeinsame-normdatei	2.020
8	lexvo	-4107.754	8	lexvo	2.017
9	rkb-explorer-acm	-4114.493	9	rkb-explorer-eprints	1.632
10	lcsh	-4273.558	10	rkb-explorer-dblp	1.466

http://www.inf.puc-rio.br/~grlopes/RecLAK/bayesian_ranking.html

http://www.inf.puc-rio.br/~grlopes/RecLAK/sn_based_ranking.html

Conclusions

- This paper presented a detailed analysis, based on Bayesian classifiers and on Social Network Analysis techniques, to address the dataset interlinking recommendation problem for LAK, using only metadata.
- The rank score functions are potentially useful to reduce the cost of dataset interlinking.

Thank you!