Searching and Ranking Educational Resources based on Terms Clustering

Marina A. H. Pimentel, Israel B. S'antanna, Marcos Didonet del Fabro
{marina, ibsa14, marcos.ddf}@inf.ufpr.br
C3SL Labs, Informatics department, Federal University of Paraná, Curitiba, Brazil

20th ICEIS, March, 21-24th, 2018
Funchal, Madeira, Portugal

CONTEXT

- Open Educational Resources (OER)
- Digital repositories
- Searching relevant OERs

PROBLEM DEFINITION

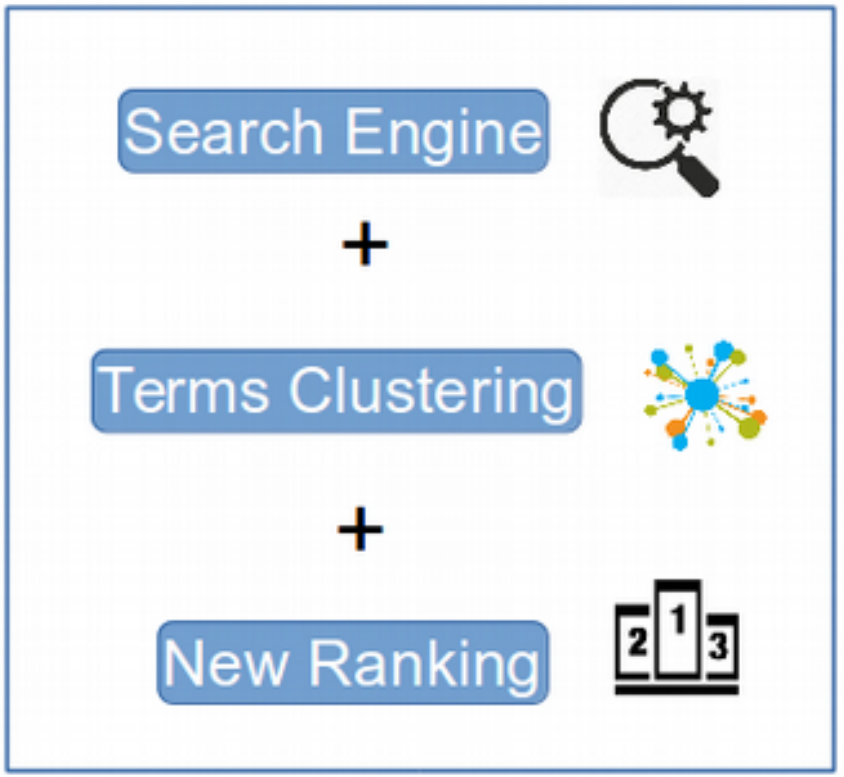
Searching OERs = exhaustive and arduous task

Main difficulties:

- restriction of syntactic search
- algorithms based only on metadata
- irrelevant results
- poorly ranked results
- user behavior

OBJECTIVE

OER Search model in digital repositories



CONTRIBUTIONS

OER Search model that improves search results:

- increasing the number of relevant OERs
- searching and ranking that considers correlated terms

OPEN EDUCATIONAL RESOURCE

OER: any educational resource openly available

Repository: collection of OERs

Metadata: information about OERs (Dublin Core Schema)

Metadata field	Value
dc.contributor.author	Moondigger
dc.date.created	2005-11-07
dc.title	Milky way 2
dc.description	Provides a close-up view of the constellation of Sagittarius
dc.subject.keyword	Astronomy
dc.subject.keyword	Constellation
dc.subject.keyword	Sagittarius
dc.subject.keyword	Space
dc.subject.keyword	Star

RELEVANCE CALCULATION

Essential to rank searching results properly.

TF-IDF: Term Frequency-Inverse Document Frequency

- relevance of a term in a document
- high value: if term occurs many times in few documents

			TF			TF-IDF		
Term	**DF**	**IDF**	***d1***	***d2***	***d3***	***d1***	***d2***	***d3***
car	18,16	1,6	27	4	24	44,5	6,6	39,6
auto	**6,72**	2	3	**33**	0	6,2	**68,6**	0
insurance	19,24	1,6	0	33	29	0	53,4	46,9
best	25,23	1,5	14	0	17	21	0	25,5

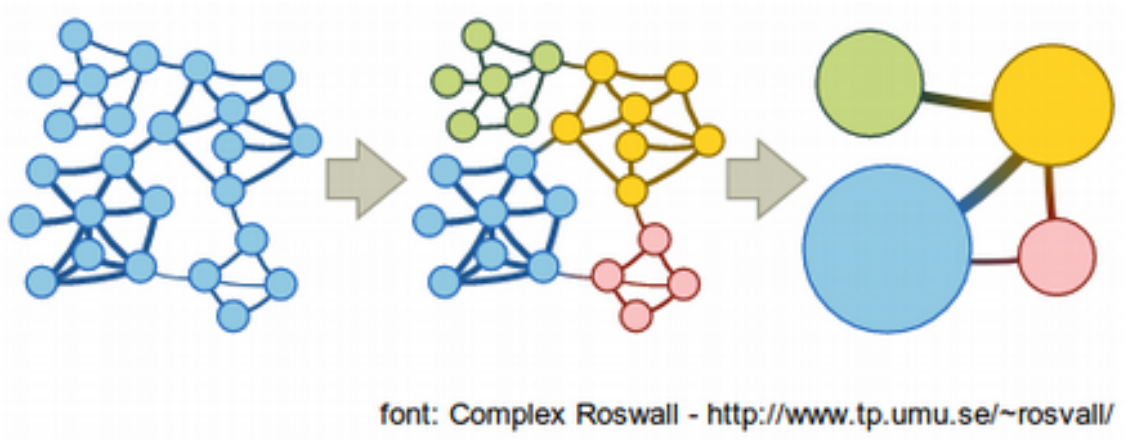
TF-IDF in a collection of 806,791 documents

DATA CLUSTERING

Groups similar objects into subsets or clusters.

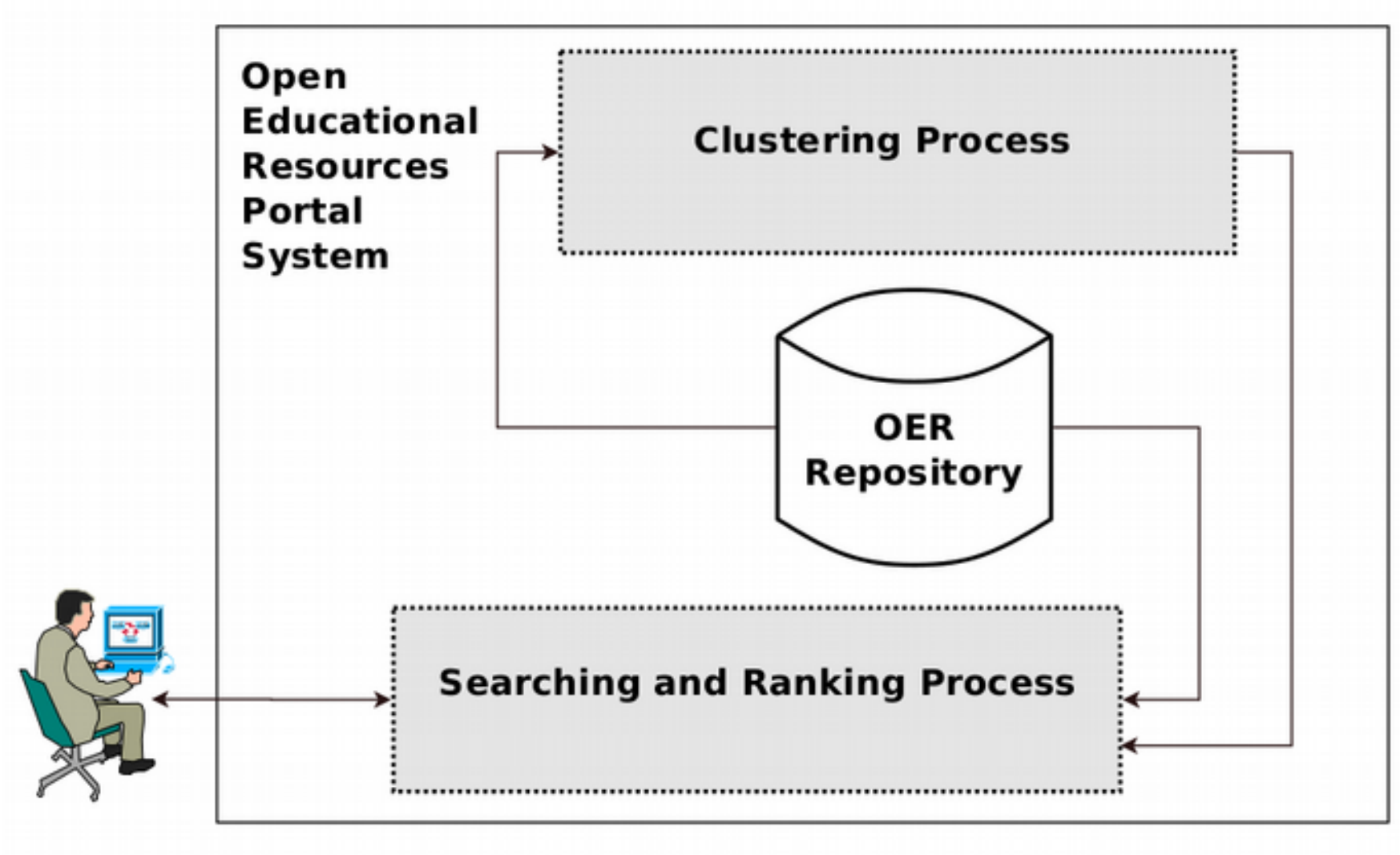
Infomap (Bohlin et al., 2014):

- community detection method
- identify strongly intraconnected modules

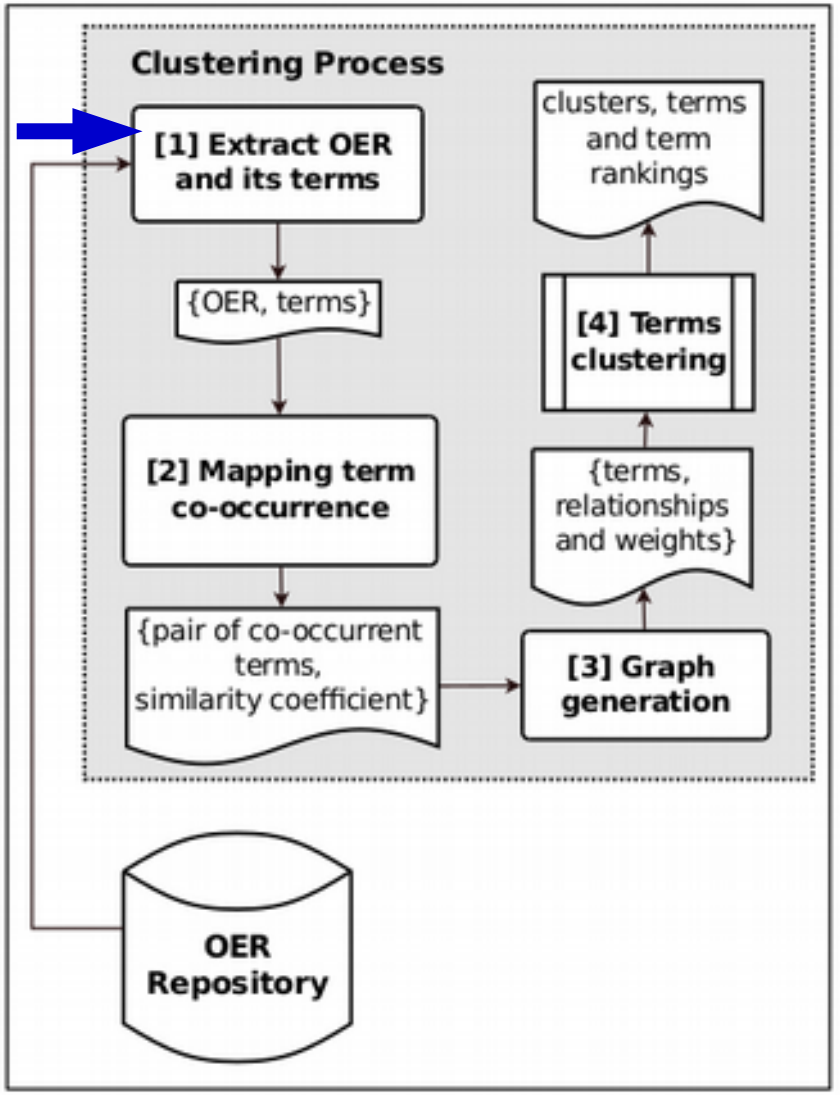


font: Complex Roswall - http://www.tp.umu.se/~rosvall/

SEARCHING AND RANKING OERs

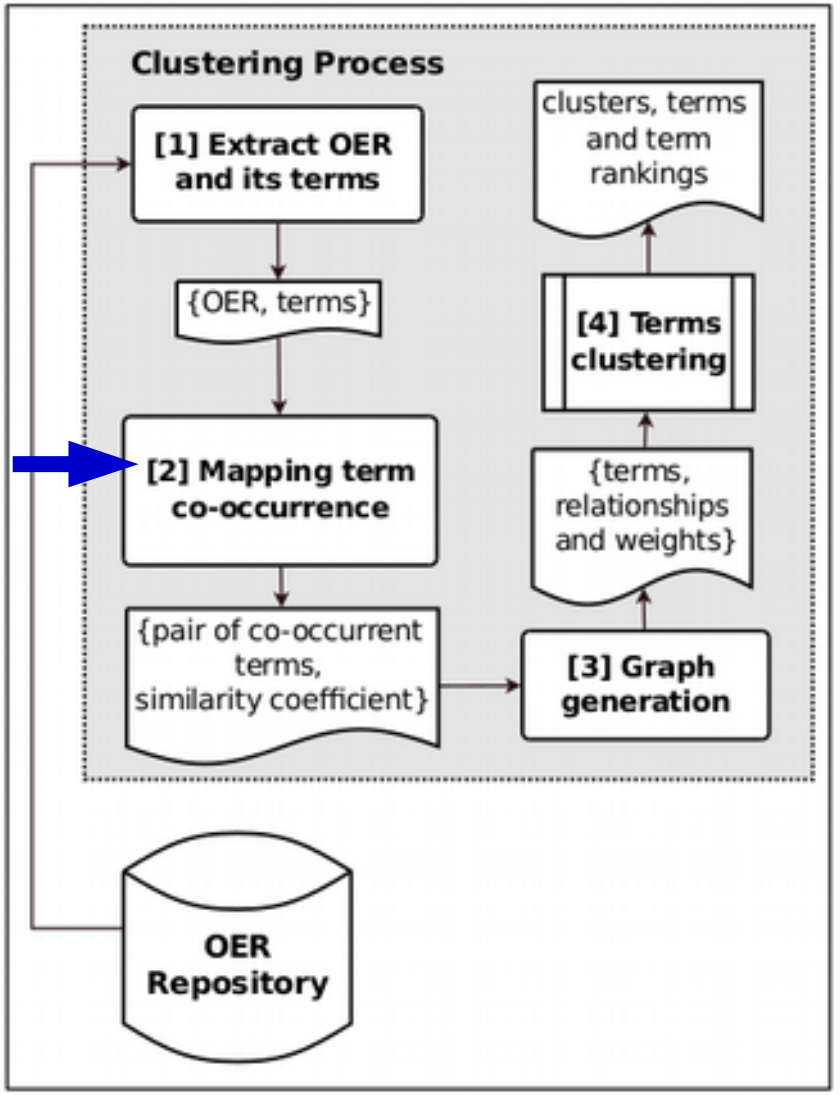


SEARCHING AND RANKING OERs



[1] S1={r1:{t10,t12}, r2: {t10,t11,t13}, r3: {t10,t11}}

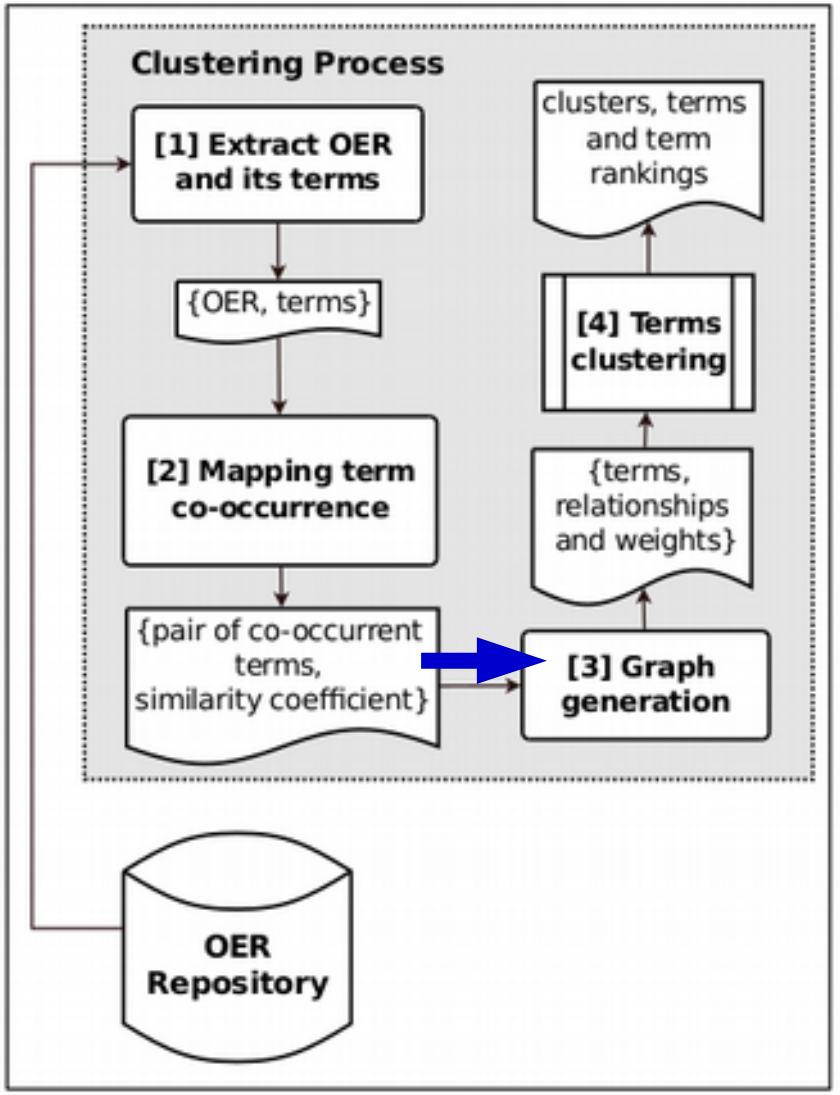
SEARCHING AND RANKING OERs



[1] S1={r1:{t10,t12}, r2: {t10,t11,t13}, r3: {t10,t11}}

[2] S2={t10: {t11: 0.82,t12: 0.58,t13: 0.58},t11: {t10: 0.82},t12: {t10: 0.58},t13: {t10: 0.58,t11: 0.71}}

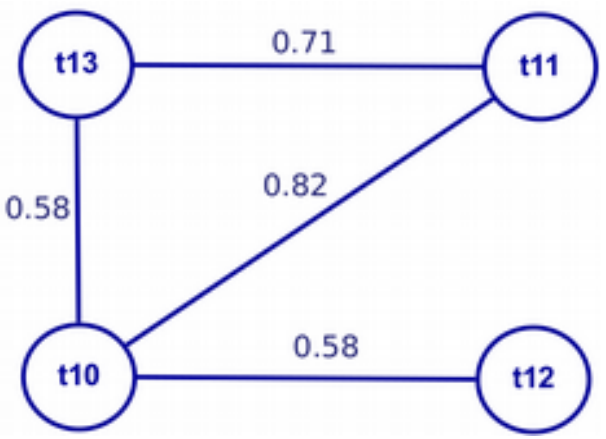
SEARCHING AND RANKING OERs



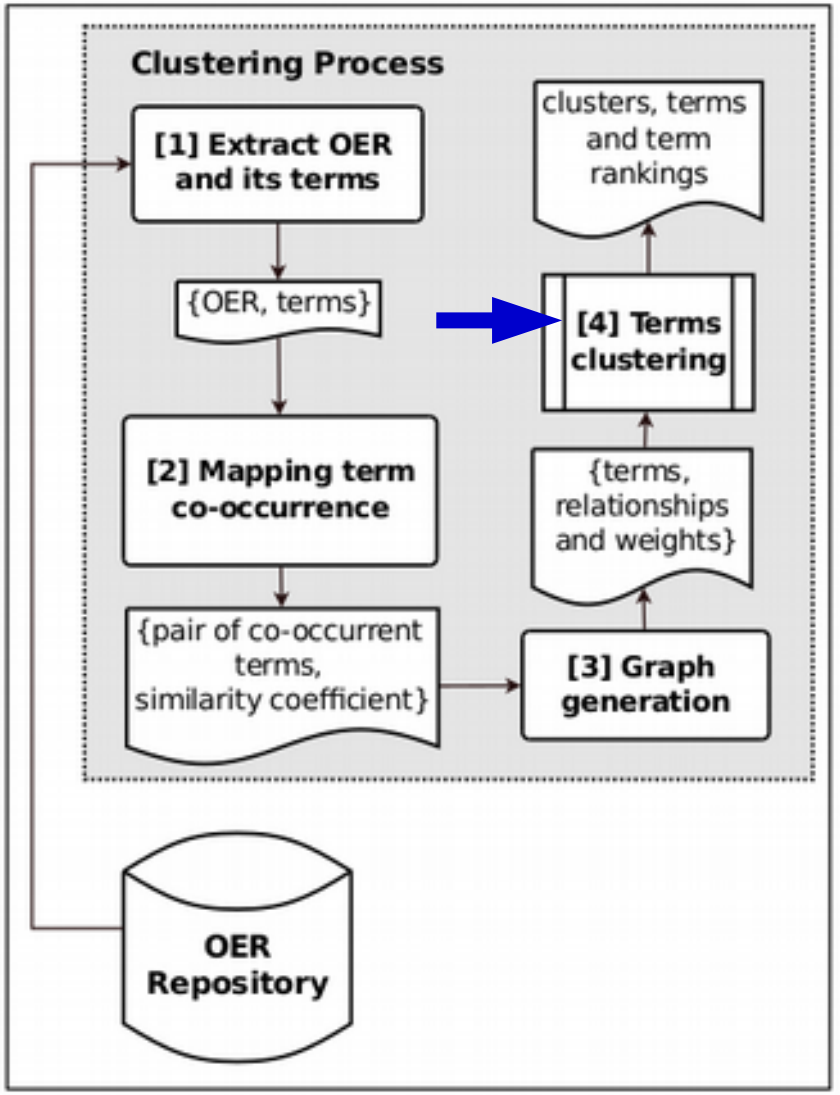
[1] S1={r1:{t10,t12}, r2: {t10,t11,t13}, r3: {t10,t11}}

[2] S2={t10: {t11: 0.82,t12: 0.58,t13: 0.58},t11: {t10: 0.82},t12: {t10: 0.58},t13: {t10: 0.58,t11: 0.71}}

[3]



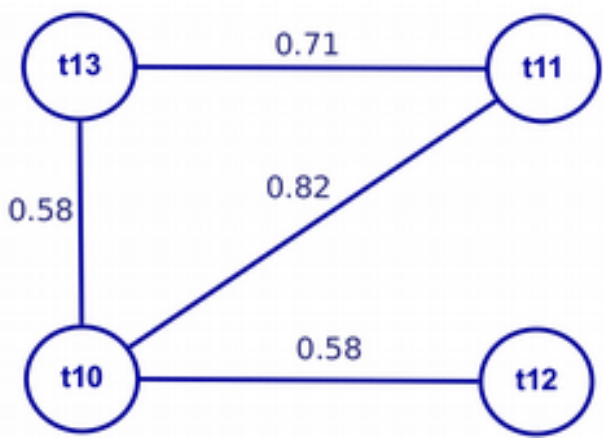
SEARCHING AND RANKING OERs



[1] S1={r1:{t10,t12}, r2: {t10,t11,t13}, r3: {t10,t11}}

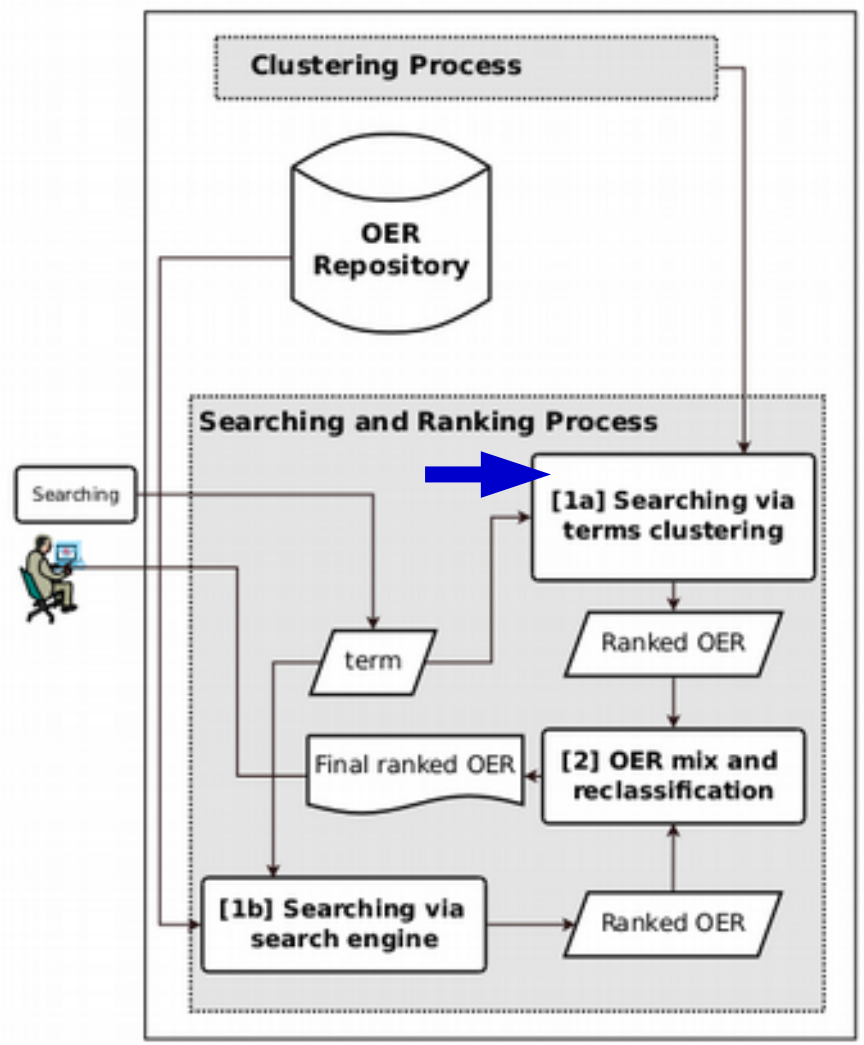
[2] S2={t10: {t11: 0.82,t12: 0.58,t13: 0.58},t11: {t10: 0.82},t12: {t10: 0.58},t13: {t10: 0.58,t11: 0.71}}

[3]



[4]

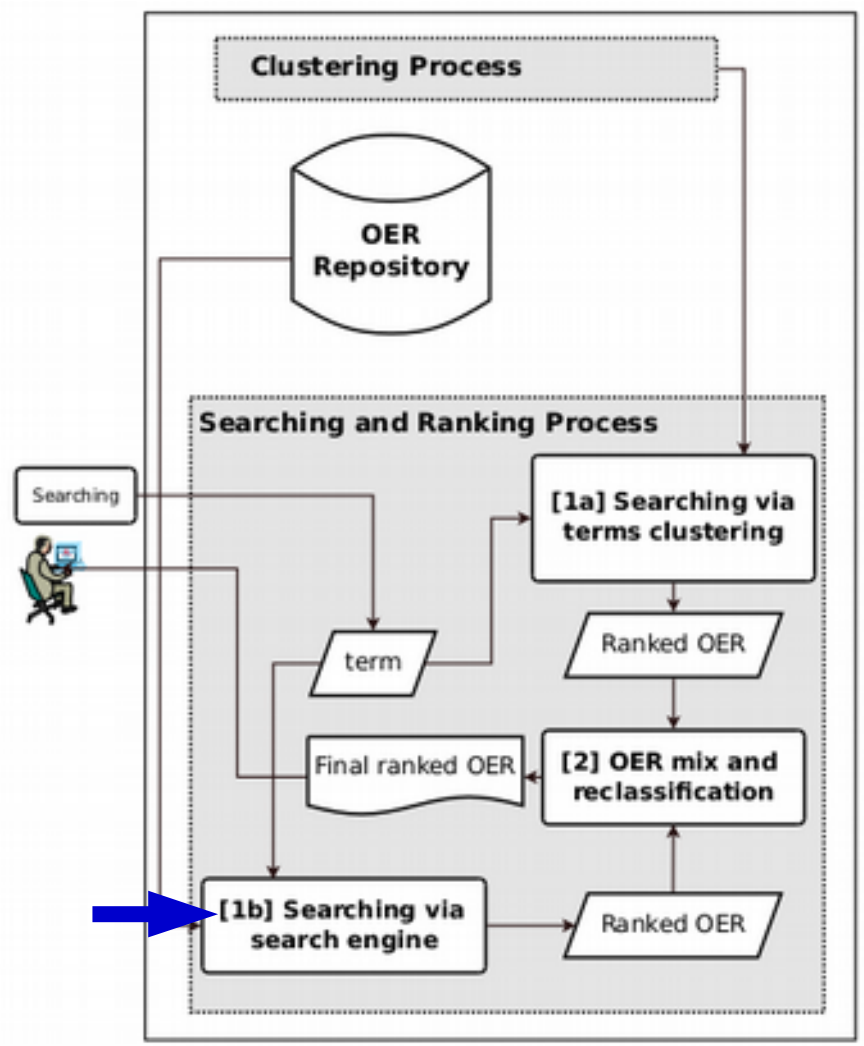
SEARCHING AND RANKING OERs



[1a]

- identifying the cluster
- recovering correlated terms
- relevance normalization of the correlated terms
- finding and ranking OERs

SEARCHING AND RANKING OERs

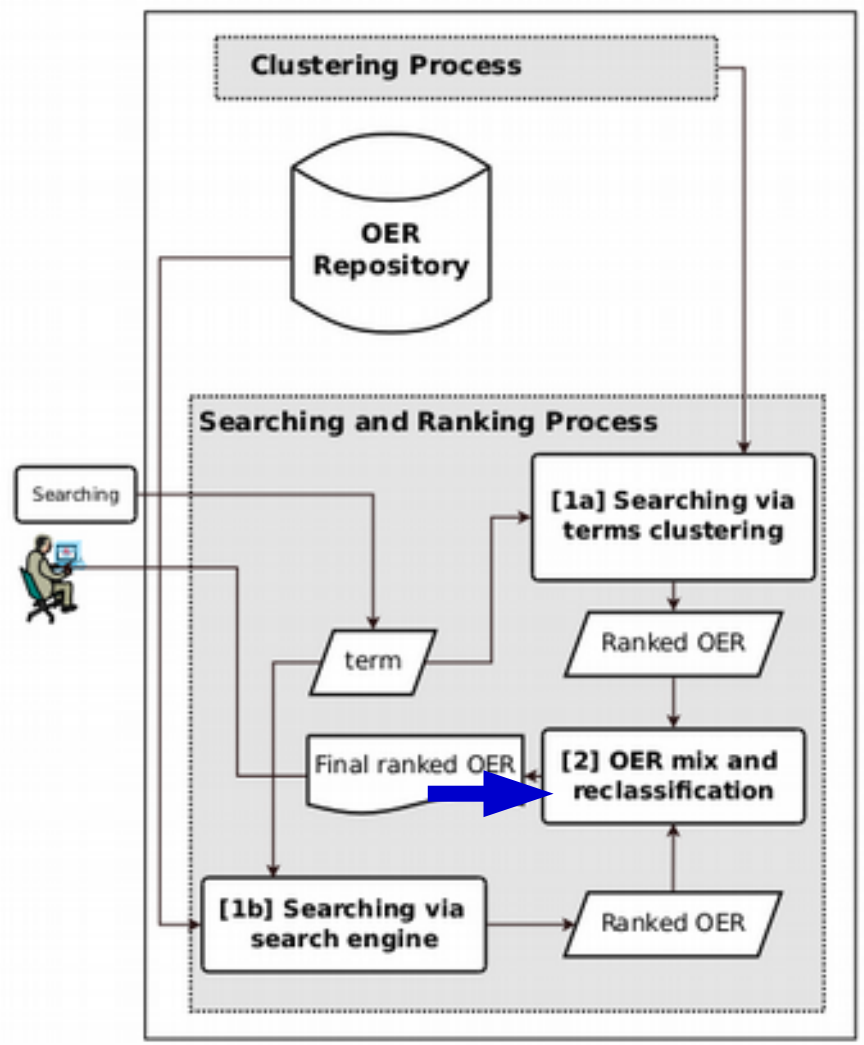


[1a] Set of ranked OERs from terms clustering

[1b]

- finding e ranking OERs via search engine
- **impulse factor (boost)**

SEARCHING AND RANKING OERs



[1a] Set of ranked OERs from terms clustering

[1b] Set of ranked OERs from search engine

[2]

- ranking values adjustment based on linear normalization (min-max)
- applying boost factor for [1a] items
- increasing OERs relevance (if in [1a] and [1b])
- returning mixed and re-ranked OERs

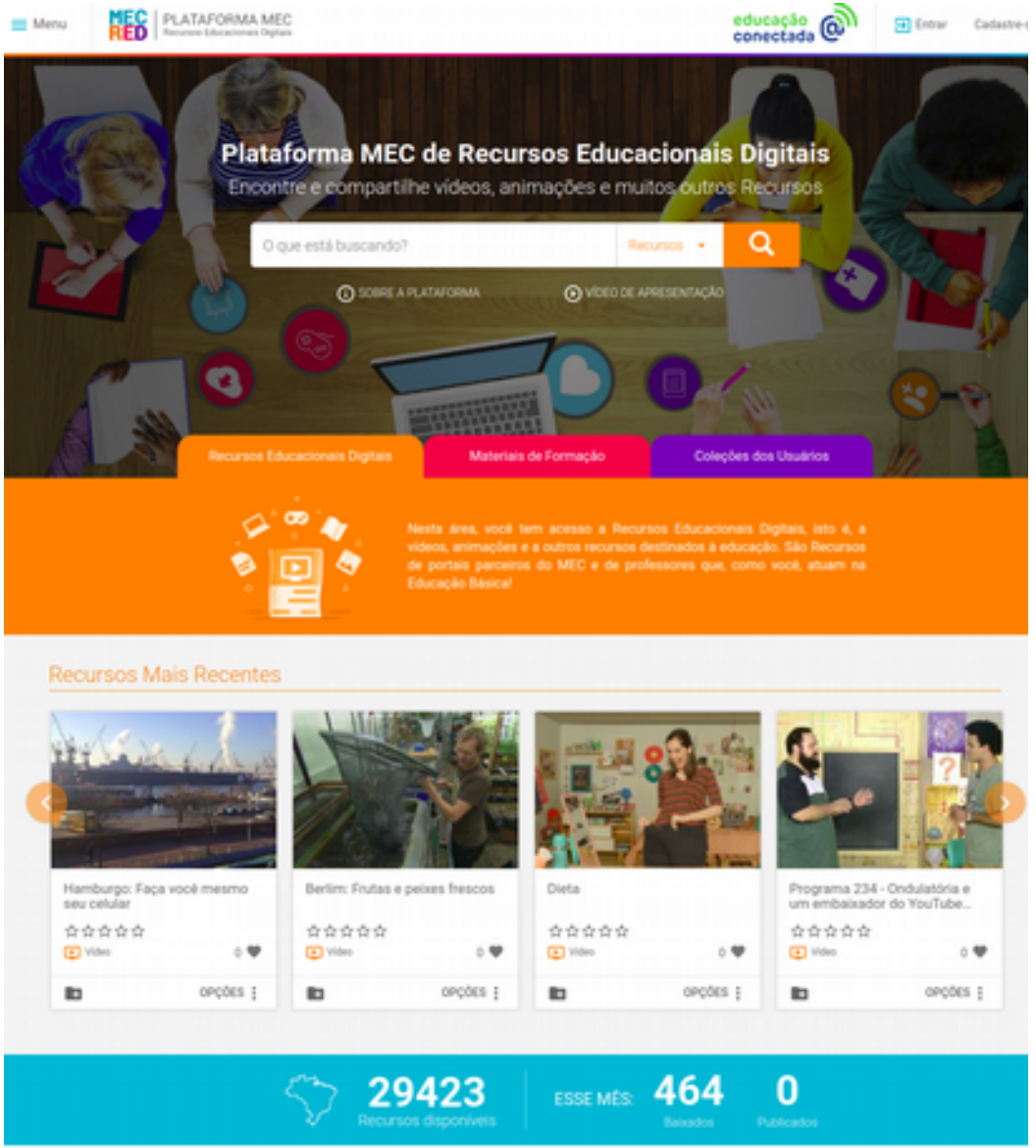
EXPERIMENTS

Open Educational Resource Platform

plataformaintegrada.mec.gov.br

Repository:

- 19,159 OERs
- 23,808 terms



EXPERIMENTS

Example of Clustered Terms

Id	Term	Original weight	Normalized weight
1	Sagittarius	2.31473e-05	1.00
2	Dwarfstar	2.42713e-05	0.99
3	Luminous star	2.42713e-05	0.98
4	Peony	3.11466e-05	0.93
5	Shine	3.11466e-05	0.93
6	W5	1.55874e-05	0.91
7	Stars movement	1.46697e-05	0.89
8	The Cartwheel Galaxy	1.18271e-05	0.84
9	Recycle	1.11963e-05	0.83
10	Protoplanetary disk	6.47752e-06	0.71
11	Hertzsprung Russell Diag.	6.13133e-06	0.70
12	Star	0.00021569	0.50

Cluster terms	Weight
DNA	1.00
RNA	1.00
Guanine	0.91
Thymine	0.82
Nucleotide	0.77
Nucleoside	0.67
Protein translation	0.62
enzymes restriction endonucleases	0.62
Homologous protein	0.61
Nucleic acid	0.59
Double helix	0.54
Capsid protein	0.50
Gravitacional forcePtolomeu Model	1.00
Retrograde movement	1.00
Position of the planets	1.00
Epicycle	0.94
Deferent	0.94
luminosity	0.87
Einstein	0.66
Greater circumference	0.64
	0.50
Corrosion	1.00
Electrochemistry	0.97
Oxide-reduction	0.95
Concentration cell	0.50

EXPERIMENTS

Searching results from

terms clustering

Search term = Sagittarius			
OER id	**OER**	**Weight**	**Term id**
2095	Stars and HR Diagram	3.1812	2,3,11,12
10667	Peony star	2.3670	4,5,12
16289	Milky way 2	1.5000	1,12
10837	W5 (Allen)	1.4114	6,12
557	W-5 Star-Forming Region	1.4114	6,12
2642	Daytime Motion of the Stars...	1.3978	12
11324	Cartwheel galaxy	1.3496	8,12
7105	Robot Astronomy...	1.3373	9,12
5482	Inner Gap in Circumstellar...	1.2147	12
11438	Space Trash	0.8373	9

search engine

Search term = Sagittarius			
OER id	**OER**	**Weight**	**Boost**
16289	Milky way 2	147.6802	10
2538	Sanitary landfill	11.1470	1
1917	Sound Almanac of Chemistry...	10.6484	1
3091	Periodical Talk - Trash	10.3336	1
17001	Sanitary landfill	9.9435	1
17393	Sanitary landfill	9.9435	1
3987	Mines without dumps	9.5448	1
6078	Slurry treatment pond	9.5448	1
15537	Mines without dumps	9.5448	1
9508	Slurry treatment	9.1596	1

EXPERIMENTS

Resulting set from OER searching based on terms clustering

Search term = Sagittarius						
	Terms clustering		Search engine		Mixed Result	
Rank	OER id	Weight	OER id	Weight	OER id	Weight
1	2095	3.18	16289	147.68	16289	13.68
2	10667	2.37	2538	11.15	2095	10.68
3	16289	1.50	1917	10.65	10667	7.03
4	10837	1.41	3091	10.33	10837	2.75
5	557	1.41	17001	9.94	557	2.75
6	2642	1.40	17393	9.94	2642	2.69
7	11324	1.35	3987	9.54	11324	2.47
8	7105	1.34	6078	9.54	7105	2.42
9	5482	1.21	15537	9.54	5482	1.87
10	11438	0.84	9508	9.16	11438	0.18

CONCLUSIONS

- Searching OERs has been an exhaustive task
- Clustered terms allows a quantitative and semantic expansion of terms
- Increase the number of relevant results
- Good results with simple ranking equations
- The implementation and experiment results show the feasibility of the approach

Searching and Ranking Educational Resources based on Terms Clustering

Marina A. H. Pimentel, Israel B. S'antanna, Marcos Didonet del Fabro
{marina, ibsa14, marcos.ddf}@inf.ufpr.br
C3SL Labs, Informatics department, Federal University of Paraná, Curitiba, Brazil

QUESTIONS?

https://plataformaintegrada.mec.gov.br/
https://gitlab.c3sl.ufpr.br/portalmec/
https://gitlab.c3sl.ufpr.br/portalmec/portalmec/tree/tag-clustering-task

REFERENCES

Aggarwal, C. C. and Reddy, C. K. (2013). Data clustering: algorithms and applications. CRC press.

Aguiar, J. J., Santos, S. I., Fechine, J. M., and Costa, E. B. (2014). Um mapeamento sistemático sobre iniciativas brasileiras em sistemas de recomendação educacionais. SBIE, 1:1123–1132.

Benediktsson, J. A., Swain, P. H., and Ersoy, O. K. (1990). Neural network approaches versus statistical methods in classification of multisource remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, 28(4):540–552.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10):P10008.

Blosseville, M.-J., Hebrail, G., Monteil, M.-G., and Penot, N. (1992). Automatic document classification: natural language processing, statistical analysis, and expert system techniques used together. In Proceedings of the 15th annual international ACM SIGIR, pages 51–58. ACM.

Bohlin, L., Edler, D., Lancichinetti, A., and Rosvall, M. (2014). Community detection and visualization of networks with the map equation framework. In Measuring Scholarly Impact, pages 3–34. Springer.

Butcher, N. (2015). A basic guide to open educational resources (OER). Commonwealth of Learning (COL);.

Coelho, G. O., Ishitani, L., and Nelson, M. A. V. (2012). Vitae: recuperação de objetos de aprendizagem baseada na web 2.0. ETD-Educação Temática Digital, 14(2):238–257.

Committee, L. T. S. et al. (2002). Ieee standard for learning object metadata. IEEE standard, 1484(1):2007–04.

REFERENCES

Costa, E., Aguiar, J., and Magalhães, J. (2013). Sistemas de recomendação de recursos educacionais: conceitos, técnicas e aplicações. In Jornada de Atualização em Informática na Educação, volume 1, pages 57–78, Campinas - SP - Brazil.

de Souza, A. B., da Silva, J. P., de Oliveira, W. C. C., Kuma, T. H., and Silveira, I. F. (2008). Recuperação semântica de objetos de aprendizagem: Uma abordagem baseada em tesauros de propósito genérico. In Brazilian Symposium on Computers in Education), volume 1, pages 603–612, Fortaleza - CE - Brazil.

Fortunato, S. (2010). Community detection in graphs. Physics reports, 486(3):75–174.

Gemmell, J., Shepitsen, A., Mobasher, B., and Burke, R. (2008). Personalization in folksonomies based on tag clustering. Intelligent techniques for web personalization & recommender systems, 12:37–48.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12):7821–7826.

Goffman, W. (1964). A searching procedure for information retrieval. Information Storage and Retrieval, 2(2):73–78.

Hassan-Montero, Y. and Herrero-Solana, V. (2006). Improving tag-clouds as visual information retrieval interfaces. In International conference on multidisciplinary information sciences and technologies, pages 25–28, Mérida - Spain.

Knautz, K., Soubusta, S., and Stock, W. G. (2010). Tag clusters as information retrieval interfaces. In System Sciences (43rd HICSS), 2010, pages 1–10, Honolulu -HI - USA. IEEE.

Lagoze, C., Lynch, C., Waters, D., Van de Sompel, H., and Hey, T. (2006). Augmenting interoperability across scholarly repositories. In Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'06), page 85, Chapel Hill - NC - USA. IEEE.

REFERENCES

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. Physical review E, 80(5):056117.

Lee, J. H., Kim, M. H., and Lee, Y. J. (1994). Ranking documents in thesaurus-based boolean retrieval systems. Information Processing & Management, 30(1):79–91.

Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C. G., and Bimbo, A. D. (2016). Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. ACM Computing Surveys (CSUR), 49(1):14–53.

Liu, R. and Niu, Z. (2014). A collaborative filtering recommendation algorithm based on tag clustering. In Future Information Technology, pages 177–183. Springer, Zhangjiajie - China.

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). Introduction to information retrieval, volume 1. Cambridge university press Cambridge, England.

Patrocinio, M. and Ishitani, L. (2009). Associação de recursos semânticos para a anotação de objetos de aprendizagem. In Brazilian Symposium on Computers in Education-SBIE), volume 1, Florianópolis-Brazil.

Pontes, W. L., França, R. M., Costa, A. P. M., and Behar, P. (2014). Filtragens de recomendação de objetos de aprendizagem: uma revisão sistemática do cbie. In Brazilian Symposium on Computers in Educatio), volume 25, pages 549–558, Dourados - MS - Brazil.

Rafailidis, D. and Daras, P. (2013). The tfc model: Tensor factorization and tag clustering for item recommendation in social tagging systems. IEEE Transactions on SMC: Systems, 43(3):673–688.

REFERENCES

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning, volume 242, pages 133–142.

Saoud, Z. and Kechid, S. (2016). Integrating social profile to improve the source selection and the result merging process in distributed information retrieval. Information Sciences, 336:115–128.

Shepitsen, A., Gemmell, J., Mobasher, B., and Burke, R. (2008). Personalized recommendation in social tagging systems using hierarchical clustering. In In proc. of 2008 ACM RecSys, pages 259–266, Lausanne - Switzerland. ACM.

Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. (1999). Analysis of a very large web search engine query log. In ACM SIGIR Forum, volume 33, pages 6–12. ACM.

Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., and Wu, T. (2009). Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In 12th EDBT, pages 565–576, Saint-Petersburg - Russian Federation. ACM.

Thet, T. T., Na, J.-C., and Khoo, C. S. (2007). Filtering product reviews from web search results. In Proceedings of the 2007 ACM symposium on Document engineering, pages 196–198. ACM.