Automatic Metadata Generation with Particle Swarms

Marko A. Rodriguez†, Johan Bollen*, Herbert Van de Sompel*

† CCS-3: Modeling, Algorithms, and Informatics, * STB-RL: Digital Library Research and Prototyping

Los Alamos National Laboratory

Abstract

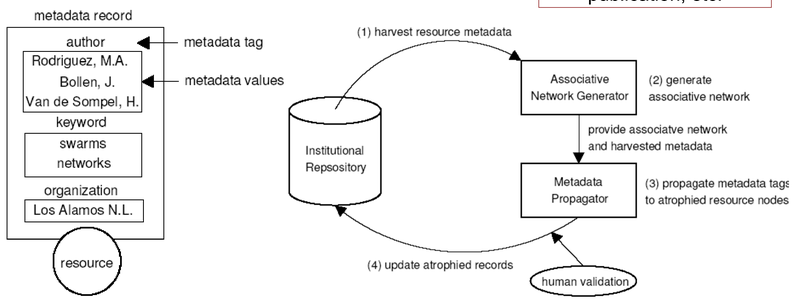
Many repositories are burdened by resources that have an incomplete metadata record. With some institutional repositories storing hundreds of millions of resources, it is extremely costly to manually generate resource metadata. Therefore, automatic metadata generation is a topic of interest to the digital library community. The automatic metadata generation system proposed by this paper is novel in three ways: it is computationally inexpensive, does not require the raw resource, and is independent of the resource media type (i.e. audio, video, document, etc.). Using occurrence and co-occurrence network generation algorithms, an associative network of repository resources is constructed using pre-existing repository metadata. The associative network serves as the substrate which allows metadata-rich resources to supply metadata-limited resources with potentially useful metadata information. This poster discusses the general framework for building associative networks, an algorithm for disseminating metadata through such networks, and a validation of the proposed system using a bibliographic dataset.

Repository Metadata and System Architecture

- For any repository resource, there exists a metadata record.
  - **Tag ⇒ Value(s)** (e.g. author ⇒ Marko A. Rodriguez)

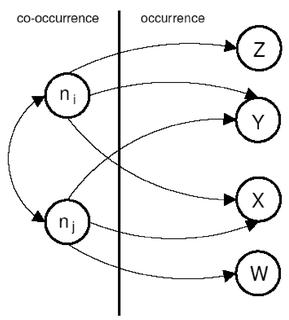
$meta(n_i, \mu)$: returns the metadata values of metadata tag μ for node n_i

Example metadata μ-types includes: authors, citations, keywords, publishing venue, organization, date of publication, etc.



Generating an Associative Network of Repository Resources

- **Occurrence networks**: resource A is connected resource B if their exists a direct reference from A to B (e.g. a citation network)
- **Co-Occurrence networks**: resource A is connected resource B if both resource A and resource B share the same metadata (e.g. a co-citation network)



- Occurrence network generation
  Running time of $O(N)$
- Co-Occurrence network generation
  Running time of $O\left(\frac{N^2-N}{2}\right)$

Occurrence Network edge weights:

$$w_{i,j} = \frac{1}{|meta(n_i,\mu)|} : n_j \in meta(n_i,\mu)$$

Co-Occurrence network edge weights:

$$co(n_i, n_j, \mu) = meta(n_i,\mu) \cap meta(n_j,\mu)$$

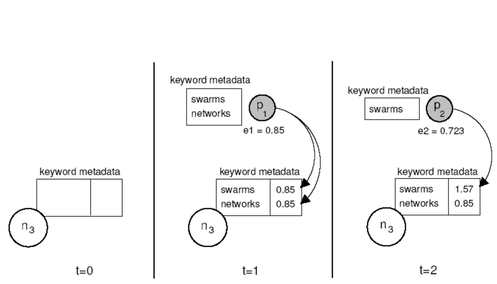
so that

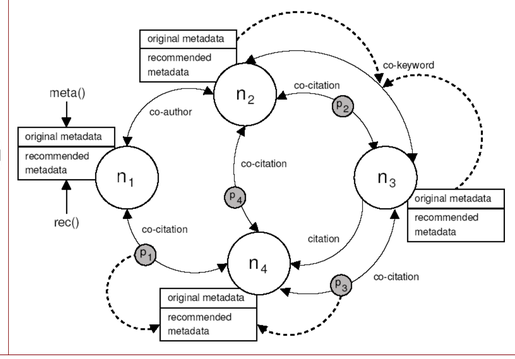
$$w_{i,j,co\mu} = \frac{|co(n_i,n_j,\mu)|}{[|meta(n_i,\mu)|+|meta(n_i,\mu)|]-|co(n_i,n_j,\mu)|}$$

All outgoing edge weights of a node are normalized to create a probability distribution for stochastic particle propagation

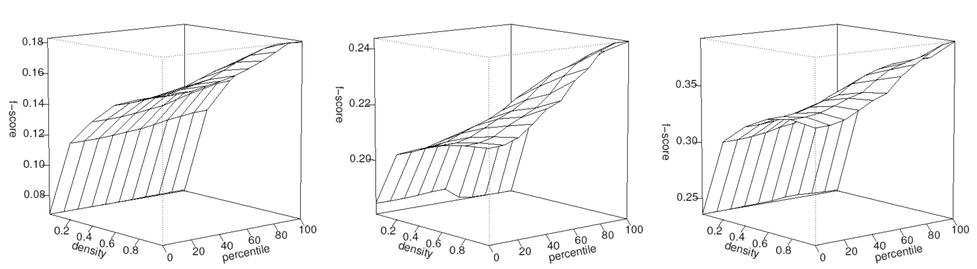
Propagating Swarms of Metadata Particles

- A particle is an indivisible entity that starts at a particular home node (resource node) and moves through the network is a stochastic manner.
  - **energy**: each step of the way the particle loses an amount of energy given by δ $e_i(t+1) = (1-\delta)e_i(t)$
  - **metadata**: a particle contains the metadata of its home node $meta(p_i,\mu) = meta(n_i,\mu) : \forall \mu$
- Metadata particles propagate over network edges in order to recommend metadata to other resources in the network. This process continues until all particle energy has decayed to 0.0.

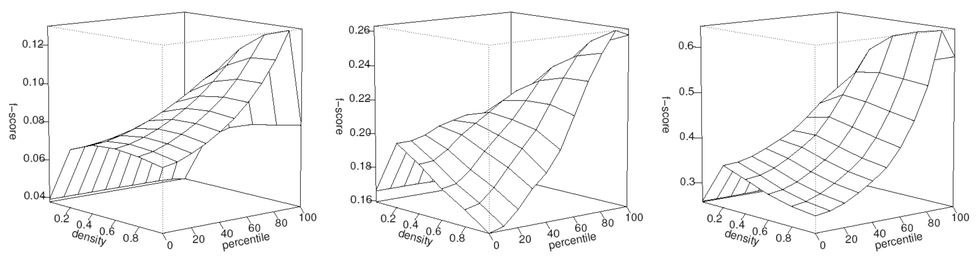




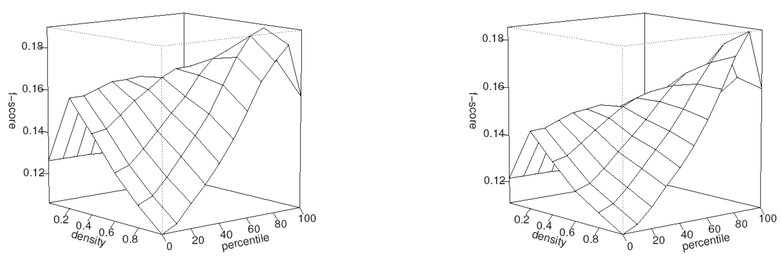
Automatic Metadata Generation Algorithm Results



Citation network propagating author, journal, and keyword metadata



Co-Authorship network propagating citation, journal, and organization metadata



Co-Keyword network and **Co-Organization** network propagating journal metadata

Validating the Metadata Generation Algorithm

Validation parameters are **Density** and **Percentile**.

- **Density**: Destroy a certain percentage of the existing metadata in the repository (1%-99%).
- **Percentile**: Allow only a certain percentile of metadata values to be accepted as valid recommendations (0%-100%). Based on metadata energy.

Validation metrics are **Precision** and **Recall**.

- **Precision**: of the recommended metadata values, what percentage are valid.
- **Recall**: of the recommended metadata values, what percentage previously existed.

$$\Pr(n_i,\mu) = \frac{|meta(n_i,\mu) \cap rec(n_i,\mu)|}{|rec(n_i,\mu)|}$$

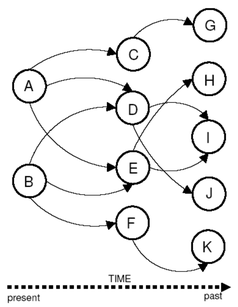
$$\mathrm{Re}(n_i,\mu) = \frac{|meta(n_i,\mu) \cap rec(n_i,\mu)|}{|meta(n_i,\mu)|}$$

$$Fscore = \frac{2 \cdot \Pr(n_i,\mu) \cdot \mathrm{Re}(n_i,\mu)}{\Pr(n_i,\mu) + \mathrm{Re}(n_i,\mu)}$$

Results Discussion

Notice how the f-score plots for the citation (occurrence) network have a different geometry than the co-occurrence network f-score plots.

A citation network isn't symmetric, therefore there is a chance that a particle will reach a dead end. When a particle reaches a dead end, it no longer recommends metadata.



Furthermore, citations are in a hierarchy with more recent publications being at the top of the hierarchy (i.e. manuscripts can not cite forward in time).

Therefore, particles trickle down the hierarchy via a single, non-recurrent path from top to bottom (similar to a 'plinko ball'). The lack of recurrence tends to produce high precision with low recall. High precision and low recall is exactly what a low percentile produces. Therefore, since the topology of the citation network yields the same effect, the effect of percentile → 0.0 isn't as pronounced.

References

[1] Collins, A. and Loftus, E. 1975. A spreading activation theory of semantic processing. *Psychological Review* 82, 407–428.

[2] Crestani, F. and Lee, P. L. 2000. Searching the web by constrained spreading activation. *Information Processing and Management* 36, 4, 585–605.

[3] Greenburg, J. 2004. Metadata extraction and harvesting: A comparison of two automatic metadata generation applications. *Journal of Internet Cataloging* 6, 4, 59–82.

[4] Han, H., Giles, C. L., Manavoglu, E., and Zha, H. 2003. Automatic document metadata extraction using support vector machines. In *Proceedings of the Joint Conference on Digital Libraries* JCDL'03. ACM, Huston, TX.

[5] Mao, S., Kim, J. W., and Thoma, G. R. 2004. A dynamic feature generation system for automated metadata extraction in preservation of digital materials. In *First International Workshopon Document Image Analysis for Libraries* DIAL'04. IEEE.

[6] Naaman, M., Yeh, R. B., Garcia-Molina, H., and Paepcke, A. 2005. Leveraging context to resolve identity in photo albums. In *Proceedings of the 5th Joint Conference on Digital Libraries* JCDL '05. Denver, CO.

[7] Rodriguez, M. A. and Bollen, J. 2005. Simulating network influence algorithms using particleswarms:Pagerank and pagerank-priors. [*submitted*].

[8] Yang, H.-C. and Lee, C.-H. 2005. Automatic metadata generation for web pages using a text mining approach. In *International Workshop on Challenges in Web Information Retrieval andIntegration*. IEEE, 186–194.