Extracting Hierarchical Topic Models from the Web for Improving Digital Archive Access

Gregory Grefenstette
Inria Saclay/TAO, Rue Noetzlin - Bât 660
91190 Gif sur Yvette, France
gregory.grefenstette@inria.fr

Lawrence Muchemi
Inria Saclay/TAO, Rue Noetzlin - Bât 660
91190 Gif sur Yvette, France
lawrence.githiari@inria.fr

ABSTRACT

Topic models provide a weighted list of terms specific to a given domain. For example, the terminology for painting, as a hobby, might include specific tools user in painting such brush, easel, canvas, as well as more specific terms such a common oil colors: deep aquamarine, cerulean blue, zinc white. For clothing, a topic model should include words such as shoes, boots, socks, skirt, hats, as well as more specific terms such as tennis shoes, cocktail dress, and specific brands of shoes, hats, shirts, etc. In addition to containing the characteristic terms of a topic, a topic model also contains the relative frequency of each term's use in the topic text. This frequency is useful in information retrieval settings; when a large number of results are returned for a query, they can be ordered by pertinence using the relative frequency of domain words to rank the responses. Providing a hierarchic topic model also allows an information retrieval application to create facets (Tunkelang, 2009), or categories appearing the result sets, with which the user can filter results, as on an online shopping site.

One problem for many information retrieval platforms in digital humanity archives is the lack of topic models, other than those already foreseen and implemented when the archive was first digitized. A researcher wishing to look at a collection or archive from a new angle has no means of exploiting a new topic model corresponding to his or her axis of research. This obstacle has two causes: (1) technologically, the platform has to allow a re-annotation of the underlying archive with a new topic model. This technological problem is solvable by implementing a suite of natural language processing tools that can access the description of the textual description of elements in the archive, and identify there terms from a new topic model. For example, the commonly used information retrieval platform Lucene (Grainger, 2014) allows the administrator to add new facet annotations to existing documents. A second, more difficult problem is (2) building a new topic model. When done manually, this is a time-consuming task, with no assurance of being complete or adequate, unless great expense is outlayed, as is the case for MeSH, a medical subject heading taxonomy (Coletti and Bleich, 2001), for which regular monthly meetings are held for maintaining and updating the terminology. For subjects less important for society, few such ontological resources exist. When topic models are created automatically they can homogenize existing terminology (Newman *et al,* 2007) but often result in noise (Steyvers, *et al.,* 2004) that may seem excessive to some archivists.

Here we will present work we have been doing to produce clean taxonomies for ad-hoc subjects, based on directed crawling and language modeling, comparing the terminology of web pages that are in-topic to a much larger collection of off-topic text. We have had some success in producing cleaner taxonomies (Grefenstette, 2015) [first place in the taxonomy creation task] and believe that that the same techniques can be applied to any domain. We are currently working on producing a wide range of taxonomies for annotation of personal data, found in user-generated text (emails, facebook post, twitter feeds).

In this talk we will present our natural language processing techniques and describe how they can applied to ad-hoc topics, creating a weighted topic model that can be used for annotating text with new facets.

References

Coletti, Margaret H., and Howard L. Bleich. "Medical subject headings used to search the biomedical literature." Journal of the American Medical Informatics Association 8.4: 317-323. 2001.

Grainger, Trey, Timothy Potter, and Yonik Seeley. *Solr in action*. Manning, 2014.

Grefenstette, Gregory. "INRIASAC: Simple Hypernym Extraction Methods". In Proceedings of Ninth International Workshop on Semantic Evaluation (SemEval 2015), Denver, 2005.

Newman, David, Kat Hagedorn, Chaitanya Chemudugunta, and Padhraic Smyth. "Subject metadata enrichment using statistical topic models." 7th ACM/IEEE-CS joint conference on Digital libraries, pp. 366-375. ACM, 2007.

Steyvers, Mark, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. "Probabilistic author-topic models for information discovery." In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 306-315. ACM, 2004.

Tunkelang, Daniel. "Faceted search." Synthesis lectures on information concepts, retrieval, and services 1.1: 1-80. 2009.