Exploring Large Document Collections using Statistical Topic Models

KDD-2006 Demo Session

David Newman*
Arthur Asuncion
Chaitanya Chemudugunta
Vasanth Kumar
Padhraic Smyth
Mark Steyvers
University of California, Irvine
{newman,..., pjsmyth}@uci.edu

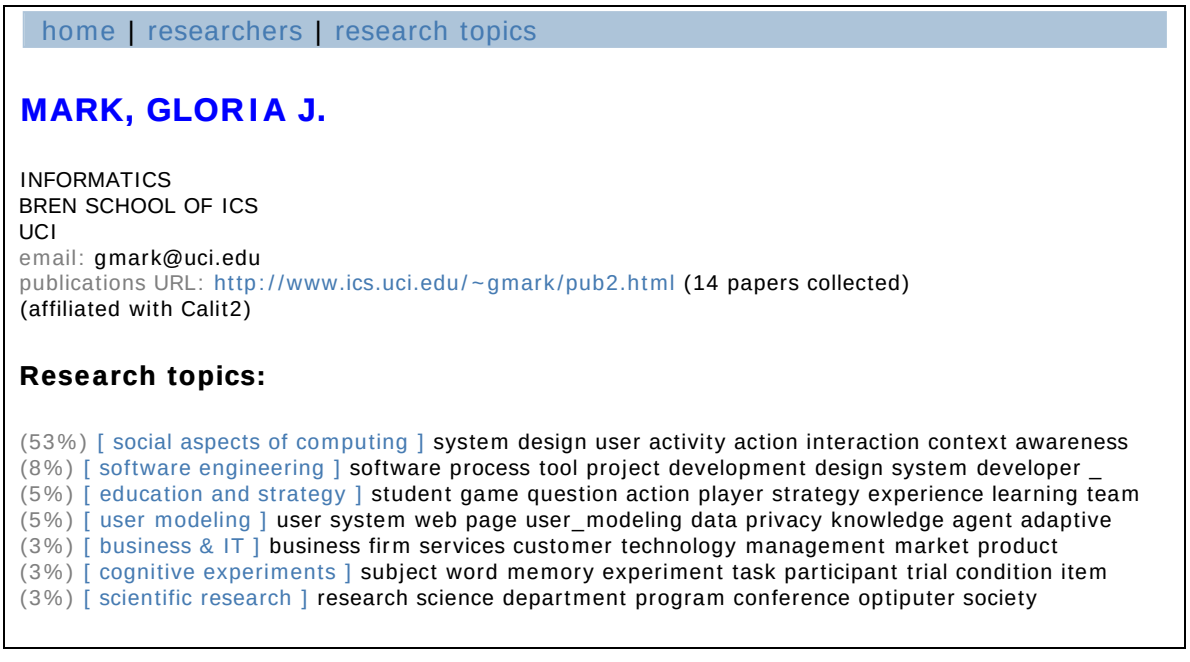
Summary

We will demonstrate the topic model, a recent unsupervised learning technique that uses a statistical model to discover topics in a large collection of text documents. The first demonstration illustrates how the topic model automatically learns about the spectrum of research conducted by faculty members at UC Irvine and UC San Diego, how to topically characterize each researcher's interests, and how to find researchers with similar interests – all in a completely unsupervised fashion. The second demonstration illustrates how medical researchers may use topic modeling to find new connections between genes and brain regions based on a large collection of articles on schizophrenia.

Demonstrations

Demo I: Calit2 browser of Researchers and Research at UCSD and UCI
http://yarra.calit2.uci.edu/calit2/

The purpose of this browser is to connect researchers across two University of California campuses. After automatically collecting 12,000 publications from 460 UCSD and UCI faculty, we used our probabilistic topic model to characterize the nature of each researcher's work and find researchers with similar interests.



Related researchers (UCSD,UCI) :

(1.0) DOURISH, JAMES P.
(0.6) OBSTFELD, DAVID M.
(0.6) REDMILES, DAVID F.
(0.5) KIRSH, DAVID J.
(0.5) HOLLAN, JIM
(0.4) GRISWOLD, WILLIAM G
(0.4) NORMAN, DONALD A.
(0.3) LOPES, CRISTINA V.
(0.3) SIM, SUSAN E.
(0.3) GIBSON, CRISTINA B.
(0.3) JAIN, RAMESH CHANDRA

Figure 1. Screen-shot from the Calit2 browser. On this page we see one particular researcher (Gloria Mark), starting with her affiliation and contact information. Below that we see the list of topics that Dr. Mark is interested in – based on the articles automatically harvested from her web page. Below that we see a list of researchers at UCSD (in green) and UCI (in blue) that closely match Dr. Mark's range of research interests. Clicking on any research topic will take the user to a page that lists the most prolific researchers in that particular research topic. The research topics were not *a priori* defined, but learned by the topic model.

Demo II: Browser of Articles, Genes and Brain Regions related to Schizophrenia

http://yarra.calit2.uci.edu/topic/schizo

The purpose of this browser is to help medical researchers mine published literature, and suggest connections that may not have been obvious. The collection is 40,000 PubMed abstracts that contain the word "schizophrenia", combined with a dictionary of 25,000 gene symbols and 5,000 brain regions. Of interest is whether the tool can suggest previously unknown topic-based connections between genes and brain regions that were hidden in the literature.

Gene symbol: comt

Prob(topic | gene): Probability of topic given the gene

(2.27%) - Syndromes from Abnormalities : syndrome autism schizophrenia deletion disorder 22q11 retardation features mental abnormalities chromosome vcf mild autistic adult

(1.88%) - Dopamine : system dopamine dopaminergic function neurotransmitter mechanism central interaction activity brain action hypothesis neurotransmission involved nt

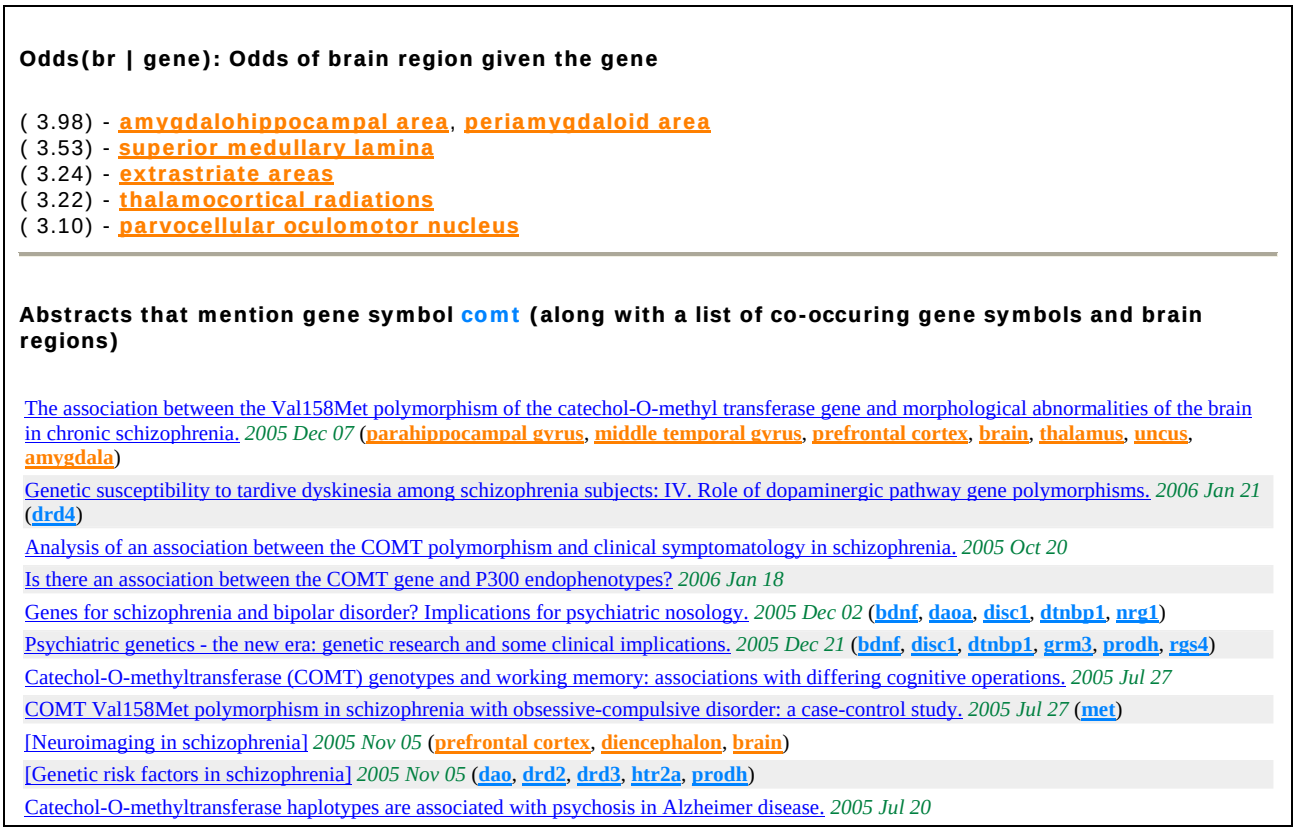
(1.66%) - Activity : activity stimulation activities enzyme low motor increased schizophrenic reduced rtm related studied normal decreased significant

(1.66%) - Cortical Regions : cortex prefrontal cortical activation frontal region functional areas temporal cingulate schizophrenia anterior brain dorsolateral parietal

(1.38%) - Performance on Neuropsychological Tests : test performance neuropsychological function patient wcst card executive sorting wisconsin cognitive frontal task measures dysfunction

Other genes with most similar topic distribution (computed via cosine distance of topic distributions)

(0.1531) - **ppp2r2b**
(0.1561) - **klhl1as**, **sca12**, **sca17**
(0.1695) - **met**
(0.1749) - **gstm1**
(0.1942) - **pml**
(0.2058) - **mthfr**
(0.2088) - **kcnn3**
(0.2095) - **il9r**



Odds(br | gene): Odds of brain region given the gene

(3.98) - **amygdalohippocampal area**, **periamygdaloid area**
(3.53) - **superior medullary lamina**
(3.24) - **extrastriate areas**
(3.22) - **thalamocortical radiations**
(3.10) - **parvocellular oculomotor nucleus**

Abstracts that mention gene symbol comt (along with a list of co-occuring gene symbols and brain regions)

The association between the Val158Met polymorphism of the catechol-O-methyl transferase gene and morphological abnormalities of the brain in chronic schizophrenia. *2005 Dec 07* (**parahippocampal gyrus**, **middle temporal gyrus**, **prefrontal cortex**, **brain**, **thalamus**, **uncus**, **amygdala**)

Genetic susceptibility to tardive dyskinesia among schizophrenia subjects: IV. Role of dopaminergic pathway gene polymorphisms. *2006 Jan 21* (**drd4**)

Analysis of an association between the COMT polymorphism and clinical symptomatology in schizophrenia. *2005 Oct 20*

Is there an association between the COMT gene and P300 endophenotypes? *2006 Jan 18*

Genes for schizophrenia and bipolar disorder? Implications for psychiatric nosology. *2005 Dec 02* (**bdnf**, **daoa**, **disc1**, **dtnbp1**, **nrg1**)

Psychiatric genetics - the new era: genetic research and some clinical implications. *2005 Dec 21* (**bdnf**, **disc1**, **dtnbp1**, **grm3**, **prodh**, **rgs4**)

Catechol-O-methyltransferase (COMT) genotypes and working memory: associations with differing cognitive operations. *2005 Jul 27*

COMT Val158Met polymorphism in schizophrenia with obsessive-compulsive disorder: a case-control study. *2005 Jul 27* (**met**)

[Neuroimaging in schizophrenia] *2005 Nov 05* (**prefrontal cortex**, **diencephalon**, **brain**)

[Genetic risk factors in schizophrenia] *2005 Nov 05* (**dao**, **drd2**, **drd3**, **htr2a**, **prodh**)

Catechol-O-methyltransferase haplotypes are associated with psychosis in Alzheimer disease. *2005 Jul 20*

Figure 2. Screen-shot from the Schizophrenia browser. Users interacting with this browser can navigate between topics, genes and brain regions. No matter which category is being viewed, the browser shows related elements from the remaining two categories, and a list of relevant documents. On this page we see one particular gene (comt), the most likely topics associated with that gene, and the most likely brain regions given that gene. Finally, a list of abstracts that mention the gene is displayed, along with the brain regions mentioned in each article. Clicking on any topic or brain region will display a similar page, while clicking on any article will pop up a window from PubMed.

Technical Specification

Our "topic browsers" allow users to interact and navigate among documents, topics, and words via a standard Web browser. The topic browsers display pre-computed data, i.e., the unsupervised learning of topics is carried out offline a priori and the browser provides an interactive framework for exploring the results. Prior to the offline learning of topic models, after preprocessing a collection of documents (using Perl scripts), and identifying entities, e.g. genes and brain regions (again using Perl scripts), the topic model is learned using a Gibbs sampling algorithm (implemented in C++ code). The learned topic model parameters are then post-processed to compute various conditional probability tables (such as the probability distribution over words given a topic) and other related quantities (using Matlab scripts). This data is then loaded into a MySQL database. The browsers display html produced by PHP scripts, retrieving data from the MySQL database in response to user navigation (clicks and search queries) in real-time.

URLs

Calit2 Browser of Researchers and Research at UCSD and UCI
http://yarra.calit2.uci.edu/calit2/

Browser of Articles, Genes and Brain Regions Related to Schizophrenia
http://yarra.calit2.uci.edu/topic/schizo

References

Griffiths, T., Steyvers, M. Finding Scientific Topics. National Academy of Sciences, 101. 5228-5235 (2004).

Blei, D., Ng, A., Jordan, M.J., Latent Dirichlet Allocation, Journal of Machine Learning Research, 1. 993-1022 (2003).

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. Analyzing entities and topics in news articles using statistical topic models. In: Springer Lecture Notes in Computer Science (LNCS) series -- IEEE International Conference on Intelligence and Security Informatics (2006).

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. Probabilistic Author-Topic Models for Information Discovery. The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2004).