Toward Discovering Potential Data Mining Applications in Literary Criticism

Bei YU
John UNSWORTH
University of Illinois at Urbana - Champaign

1. Introduction

Over the past decade text mining techniques have been used for knowledge discovery in many domains, such as web documents, news articles, biomedical literature, etc. In the literary study domain, some data mining applications have emerged, among which document categorization may be the most successful example (Meunier 2005). But the overall progress of computer assisted literary study is not significant.

The goal of this research is to discover more potential data mining applications for literary study. The basic belief underneath our research is that in order to better adapt data mining techniques to literary text, one has to grasp the unique characteristics of literary research and to leverage its uniqueness and its similarity with data mining. Buckland (Buckland, 1999) claimed that vocabulary is a central concept in information transition between domains. Comparing the vocabularies between the corpora in different domains may shed light on discovering the similarity and difference in the research activities between these domains. So we propose a 3-stage approach to map research activities between data miners and literary scholars as reflected in the vocabulary use in their research publications. Stage 1 is to investigate literary scholars' unique research activities by verb analysis and topic analysis in critical literature, and see if any available data mining techniques can be applied to assist the scholars in these activities. Stage 2 is to investigate the mainstream data mining practices and the representations of the discovered knowledge by keyword analysis in data mining literature, and see if they also appear in critical literature setting. The shared research activities and knowledge representations will suggest some research problems on which data mining experts and literary scholars can start their collaboration. The two stages are complimentary to each other rather than sequential. In the last stage, potential literary text mining problems are summarized into a list of questions, and some literary scholars are interviewed to verify if these applications are useful and which of them can be specified to be ready for text mining.

Up to date we have finished the first two stages. We will be interviewing 5-10 literary scholars between now and the conference. The results of the interviews will be included in our presentation at the conference.

2. Corpus Construction

Three corpora have been constructed for the vocabulary use analysis in stage 1 and 2. The first is the data mining corpus (named "KDD") which consists of 442 ACM SIGKDD conference paper abstracts from 2001 to 2005. The ACM SIGKDD conference has been the premier international conference on data mining. The paper titles and abstracts are extracted from the ACM Digital Portal. We do not use full text because it contains too many technical details that are not relevant to literary research.

The second is the literary criticism corpus (named "MUSE") which consists of 84 ELH Journal articles and 40 ALH articles downloaded from Project Muse, all on the subject of the 18th and 19th century British and American literature. The selection is based on the subject indexes assigned by the publisher. The plain text versions are generated by removing all the tags and quotations from the corresponding HTML versions.

The third is the New York Times subset of American National Corpus (named "ANC-NYTIMES") which consists of thousands of news articles with more than 2 million words. This "everyday English" corpus serves as a contrast group to test if the discovered similarities between the research behaviors in data mining and literary study are significant.

3. Stage 1: discovering literary scholars' unique research activities

This stage consists of three steps. Firstly, the plain text MUSE documents are part-of-speech tagged using GATE. Document frequency (DF) and term frequency (TF) serve as the basic indicators for a term's

popularity in a collection. Arbitrary DF is defined as the number of documents that contain the term. Normalized DF is defined as the percentage of the arbitrary DF in the collection (denote as "DF-pcnt"). Arbitrary TF is defined as the term's total number of occurrences in the whole collection. Normalized TF is defined as the proportion per million words (denote as "TF-ppm"). The verbs are cascade sorted by their DF and TF.

A literary scholar picked out some representative verbs (with both DF and TF between 5 and 10) in critical literature setting: "clarifies", "cleared", "Knowing", "destabilizes", "analyzing", "annotated", "juxtaposed", "evaluated", "recapitulates", "merit", "detail", "portraying", and "stemming".

Secondly, a unique MUSE verb list is generated by comparing the verbs in MUSE and ANC-NYTIMES, also cascade sorted by DF and TF. The top 10 unique verbs are "naturalizing", "narrating", "obviate", "repudiate", "Underlying", "misreading", "desiring", "privileging", "mediating", and "totalizing".

Obviously the two verb lists do not overlap at all. Actually, the representative verbs (except "recapitulates") picked out by the literary scholar turn out to be common in ANC-NYTIMES corpus too. After examining the unique MUSE verb list, two literary scholars were surprised to find many unexpected unique verbs, which means their uniqueness is beyond the scholars' awareness.

Thirdly, simple topic analysis shows that many MUSE essays are trying to build connections between writers, characters, concepts, and social and historic backgrounds. As an evidence, 56 out of 84 ELH essays and 24 out of 40 ALH essays titles contain "and" - one of the parallel structure indicator. But genre is the only topic that can be mapped directly to text mining application - document categorization.

In conclusion, literary scholars are not explicitly aware of what are the unique research activities at the vocabulary-use level. They might be able to summarize their scholarly primitives as Unsworth did in (Unsworth, 2000), but does not help computer scientist to understand the data mining needs in literary criticism.

4. Stage 2: discovering the mainstream data mining activities and the representations of discovered knowledge in KDD and MUSE corpora

This stage of analysis consists of two steps: 1) extracting keywords from KDD paper titles, identifying mainstream data mining activities and knowledge representations in data mining; and 2) comparing the DFs and TFs of the KDD keywords between KDD, MUSE, and ANC-NYTIMES corpora, identifying the keywords common in both KDD and MUSE but not in ANC-NYTIMES.

In the first step, non-stop words are extracted and stemmed (using Porter Stemmer) from paper titles and sorted only by their TF. 18 out of 102 non-stop stemmed title words with TF>5 are identified as the representative data mining keywords. The left out terms include general terms (e.g. "approach"), technical terms (e.g. "bayesian"), terms about specific data (e.g. "gene"), and terms with different meaning in MUSE (e.g. "tree").

Table 1 compares the frequencies of the 18 words between MUSE and ANC-NYTIMES. It shows that 11 data mining keywords are common in literary essays but not in news articles. Figure 1 visualizes their significant differences in TF-ppm. The 11 keywords stand for models, frameworks, patterns, sequences, associations, hierarchies, classifications, relations, correlations, similarities, and spatial relations. It's not surprising that none of these keywords can be found in MUSE essay titles. The context of the keywords extracted from KDD abstracts and MUSE full text also has little in common.

In the left 7 KDD keywords, "rule", "serial/seri" and "decis" are common in both corpora, "cluster" and "stream" are common in neither of them. Interestingly "network" and "graph(ic)" are much more common in ANC-NYTIMES. It seems literary scholars do not think much in graphic models.

In conclusion, literary scholars are actually "data miners", except that they look for different kinds of knowledge. For example, in terms of pattern discovery, literary scholars look for "narrative patterns", "marriage patterns", "patterns of plot", etc. But data miners concern pattern in a more abstract manner - "sequential patterns", "association patterns", "topological patterns", etc.

5. Stage 3: interview the literary scholars

to verify the potential literary data mining applications

In this stage we are going to interview 5-10 literary scholars to examine 1) how the scholars discover the kinds of knowledge identified in stage 2; 2) how to specify these kinds of knowledge so that computational algorithms can be designed to discover them for literary study purpose.

KDD Title Words	KDD TF	MUSE DF	MUSE TF	MUSE DF-pcnt	MUSE TF-ppm	ANC-NYTIMES DF	ANC-NYTIMES TF	ANC-NYTIMES DF-pcnt	ANC-NYTIIMES TF-ppm
cluster	52	13	17	10	22	50	56	1	24
model	40	99	438	80	562	335	597	8	254
pattern	29	49	121	40	155	167	207	4	88
Net-work	23	30	76	24	97	325	724	8	307
classif classifi	35	14 19	64	11 15	82	16 50	74	0 1	32
rule	19	81	210	65	269	708	1409	17	598
associ	15	103	479	83	614	762	1161	18	493
graph graphic	15	2 15	25	2 12	32	7 244	504	0 6	214
stream	15	19	26	15	33	103	117	2	50
serial seri	10	20 80	213	16 65	273	32 537	94	61 13	403
relat relation-ship	10	117 98	1367	94:79	1753	386 323	911	9 8	487
framework	10	38	69	31	88	25	28	1	12
correl	9	20	30	16	38	15	21	0	9
similar similarli	9	99 66	475	53 80	609	484 77	649	12 2	276
spatial	7	19	55	15	71	8	8	0	3
decis	7	57	151	46	206	683	1110	16	471
hierar-ch(i)	6	20 41	178	16 33	229	4 4	45	0 0	13
sequen(c/ti)	6	34 8	80	6 27	103	4 51	71	0 1	30

Table 1: KDD keyword frequency comparison between MUSE and ANC-NYTIMES

Note: Because of the limitation of Porter Stemmer, some words with the same stems have to be manually merged together, such as “graphs” and “graphics”. In these cases the TF-ppm can be summed up, but the DF-pcnt can not be merged, so both DF-pcnts are listed.

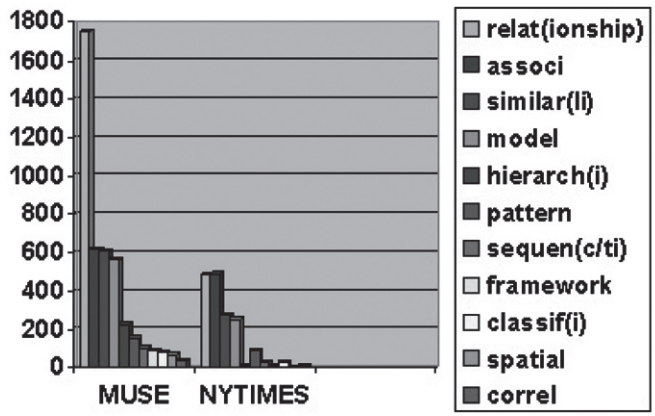

Figure 1: The frequencies (in ppm) of KDD keywords in MUSE and NYTIMES

References

Buckland, M. (1999). Vocabulary as a central concept in library and information science. In Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities, *proceedings of the Third International Conference on Conceptions of Library and Information Science*. Dubrovnik, Croatia, pp. 3–12.

Mei, Q. and Zhai, C. (2005). Discovering evolutionary theme patterns from text. *Proceedings of The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Chicago, Illinois, pp. 198-207.

Meunier, J. G., Forest, D. and Biskri, I. (2005). Classification and categorization in computer-assisted reading and text analysis. In Cohen, H. and Lefebvre, C. (eds), *Handbook on Categorization in Cognitive Science*. Elsivier.

Unsworth, J. (2000). Scholarly primitives: what methods do humanities researchers have in common, and how might our tools reflect this? *Symposium on Humanities Computing: formal methods, experimental practice*. King’s College, London. Available: http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html