Content Modeling Paradigm: An Interplay of Relationship between Author, Document, Topic, and Words

Deepak Gupta
Delhi College of Engineering,
Delhi, India

ABSTRACT

For any work of literature, a fundamental issue is to identify the individual(s) who wrote it, and conversely, to identify all of the works that belong to a given individual or to identify the individual who writes many papers on same topic or to identify the topics name that an author works on. Information extraction techniques (such as Author Name and Topic Recognition) have long been used to extract useful pieces of information from text. The types of information to be extracted are generally fixed and well defined. However in some cases, the user goal is more abstract and information types cannot be narrowly defined. For example, a reader of online user reviews typically has the goal of making a good choice and is interested to learn about the different aspects of a topic and author relation (e.g., famous author of a topic, author's papers with his research field). Some of these aspects may be known by the reader and some others may need to be discovered from the inherent text structure in a large collection. Even for the known aspects (such as "author name" and "topic"), the challenge is to recognize various hidden aspects like number of papers written by an author, his research field, popularity of an author.

In this paper, we will develop content modeling Paradigm to extract the relationship between the author, document, topic and Words as topics with identifiable word distributions across documents of various authors. We review several probabilistic graphical models (such as Latent Dirichlet Allocation) and propose a new model called content modeling paradigm which is based on frequency of the words within the document.

Index Terms — Data mining, ATP Model, TAP Model, Content modeling, supervised paradigm, unsupervised paradigm

I. INTRODUCTION

Content extraction is typically performed in the following setting: given a content type in a type system, specify all segment of manuscript which is instances of this type. A type system is similar to a database schema where we define the semantics for each field. The types of desired information is fixed and while the expression of this information can vary to a large extent in manuscript, it is often the case that many contextual clues and pattern that can be learned for these extraction remains the same. In some cases (e.g., web pages), it is even possible to take advantage of clues other than textual contents for the extraction (e.g., formatting differences for the names of the person on their personal web page vs. other contents).

1.1 Problem Statement

We focus on the special type of the problem described above where we are interested to extract the various relations between topic and author-name within the documents. We are looking for a concise answer to the questions of "what do I need to know about topic and author". Our terminology is summarized in Figure 1-1 and Figure 1-2. In the Figures Author-Name is the name of the author who wrote the papers on various topics of his field either individually or with some co-authors on his own research field or not. The papers are the author's papers which are written of its own or with collaboration with another author called co-author. Research field indicates the core topic of research of an author and we have assumed that an author always repeats his researcher topic related words in his papers. The line between author-name, papers, research field in the Figure 1-1 refers to the fact that we have to search author's papers as well as his research field with some priori knowledge i.e. author-name. The task is to recognize the mentions of each paper and research field using author-name while papers of various authors using topic-name.

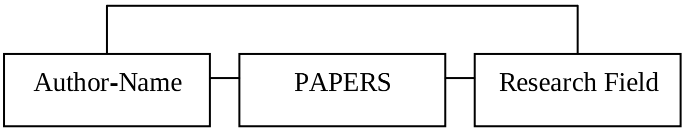


FIGURE 1-1: TERMINOLOGY OF AUTHOR-TOPIC RELATIONSHIP

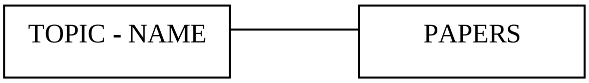


FIGURE 1-2: TERMINOLOGY OF TOPIC-AUTHOR RELATIONSHIP

Thus the above problem is further partitioned into the following objectives:

1. Can we prepare a system for extracting author-topic relation?
2. Can we develop a system for the above based on frequency of words?
3. Can this system assist the new researchers?

On the surface, this seems like a typical manuscript classification task but some features of the task make it hard for the conventional classification method:

1. The list of content is open-ended and needs to be discovered from the corpus. As was the case, described above in the example, the aspect labels often are not explicitly mentioned in the text. This is similar to problem of finding cluster labelling in unsupervised learning.
2. There is considerable variability in the contexts of the author-name we would like to extract. For example, an author named "Deepak" can be so many, so there will be much ambiguity in the author-name relation.
3. There is an issue with the name of the topic, i.e. one topic can have multiple related words for example algorithm have its related words: program, analysis, design, pseudo code, techniques etc.

1.2 Contributions

In this paper, we will use this knowledge to design our approach based on frequency of words within the document called content modelling paradigm: an interplay of relationship between author, document, topic and words. There are many probability based author-topic model are available which searches the papers on the basis of the probability of words present in the author's document, but here we have introduce another technique based on frequency through which we can search the papers according to the frequency of repeated words within the document. Probability based author-topic model have better performance compared to our frequency based model in discovering the required information and specifying the corresponding span of text in terms of accuracy and time. We have designed the system which is less complicated and it reduces the vast calculations just by increasing the total number of searches or comparisons.

To accomplish the above objectives, we proposed the following system which constitutes the contribution made:

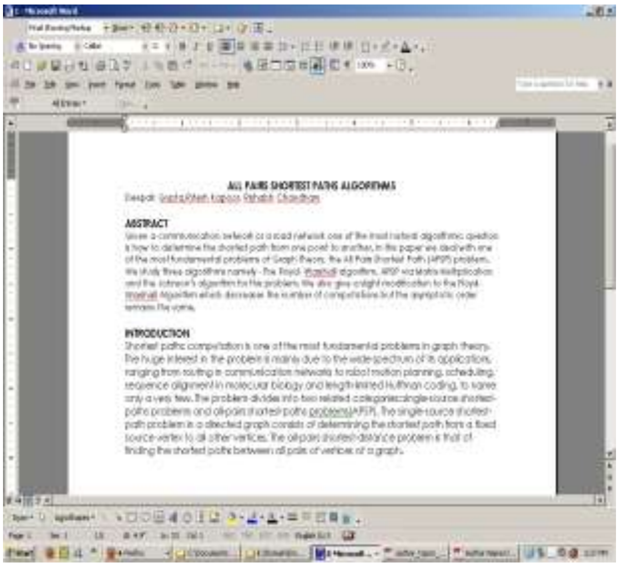
1. To accomplish the first objective, we had developed two algorithms for extracting topic-author relationship and both the models uses some documents containing some words.

2. The system proposed above is based on the concept that a research paper consists of the words maximum number of times which are related to the research field of the paper and we will use bag-of-words to extract topic, author, document and words relationship.

3. The proposed system is very useful to the new researchers as initially they don't aware of the research areas where they can work on, and they also don't know the sequencing of the papers of an author on the same field, this system assist the researcher in searching the papers they are interested to search.

1.3 Assumptions

1. Our project is doing best work when it is being installed for a web search engine, where there is a fixed document format of a research paper as shown below:



ALL PAIRS SHORTEST PATHS ALGORITHMS

Deepak Gupta,Ritesh Kapoor, Rishabh Chaudhary

ABSTRACT

Given a communication network or a road network one of the most natural algorithmic question is how to determine the shortest path from one point to another. In this paper we deal with one of the most fundamental problems of Graph theory, the All Pairs Shortest Path (APSP) problem. We study three algorithms namely - The Floyd- Warshall algorithm, APSP via Matrix Multiplication and the Johnson's algorithm for the problem. We also give a slight modification to the Floyd-Warshall Algorithm which decreases the number of computations but the asymptotic order remains the same.

INTRODUCTION

Shortest paths computation is one of the most fundamental problems in graph theory. The huge interest in the problem is mainly due to the wide spectrum of its applications, ranging from routing in communication networks to robot motion planning, scheduling, sequence alignment in molecular biology and length-limited Huffman coding, to name only a very few. The problem divides into two related categories:single-source shortest-paths problems and all-pairs shortest paths problems(APSP). The single-source shortest-path problem in a directed graph consists of determining the shortest path from a fixed source-vertex to all other vertices. The all-pairs shortest-distance problem is that of finding the shortest paths between all pairs of vertices of a graph.

Figure 1-3: Format of research paper

a. The Paper should strictly follow the above mentioned format.

b. The names of all the author and co-authors (if any) is in the second line separated by a comma (,).

2. The algorithm does not take sentence structure into account.

3. The algorithm uses simplistic statistical methods which work well with a large amount of data, but return inaccurate results when applied to small data sets.

4. Ambiguity of Author(s) is not removed.

5. Ambiguity of Topic(s) is not removed.

6. Author will use his research field word more number of times compared to the words which is not related to his research field.

1.4 Organization of paper

The rest of this paper is structured as follows: Section II summarizes the complete details design of the new approach, their algorithms, and limitations. Section III shows the basic algorithm used in the content modeling, section IV provide the implementation of the model using the content modeling algorithm. Section V shows the experimental setup and results of the proposed system and finally the paper concludes in Section VI.

II. DESIGN AND ARCHITECTURE

In this section we will elaborate the two paradigm approaches used to design such systems and discusses that which approach we had used in our system designing. This section also involves the discussion of the two models i.e. ATP and TAP models.

2.1 Types of machine Paradigms

Most disambiguation paradigms fall into one of the two machine learning paradigms:

a. supervised or
b. Unsupervised.

Supervised paradigm take as input a set of training examples consisting of pairs of articles that are categorized as either positive (author match) or negative (not author match), while unsupervised paradigm do not use categorized training examples. In general, supervised approach performs better as they are tuned specifically to the data (e.g., to determine the relative significance and interactive effects of dissimilar features such as a journal name vs. co-author vs. affiliation vs. title word). Having a adequate amount of training data is critical to the performance of any predictive model that will be extrapolated to new heretofore-unseen examples. The amount of data sufficient for training depends on the complication of the model. Bayesian unsupervised learning is used to fit the model to a document collection.

The author-topic models can be used to support a variety of interactive and exploratory queries on the set of documents and authors, including analysis of finding the authors who are most likely to write on a given topic, topic trends over time and finding the most unusual paper written by a given author.

Automatic retrieval of topics from text, via unsupervised learning paradigm, has been addressed in prior work using a number of different approaches. One general approach is to represent the high-dimensional word vectors in a lower-dimensional space. Local regions in the lower-dimensional space can then be associated with specific topics.

2.2 Content modeling Paradigm

The originality of the work described in this paper lies in the proposal of a frequency statistic model called content modelling paradigm that represents authors, document, words and topics, and the application of this model to a huge well-known document corpus in computer science. As we will show later in the paper, the model provides a general framework for discovery, exploration and query-answering in the context of the relationships of topics and author for large document collections. With the introduction of the Web and various specialized digital libraries, the automatic retrieval of useful content from manuscript has become an increasingly significant research area in data mining. In this paper we talk about a new algorithm that retrieves the topics expressed in large text document collections using bags of words and modelled how this algorithm can retrieve topic and papers of an author.

Figure 3-1 and 3-2 shows the basic flow of information discovery system and it just includes the inputs and outputs of the model.

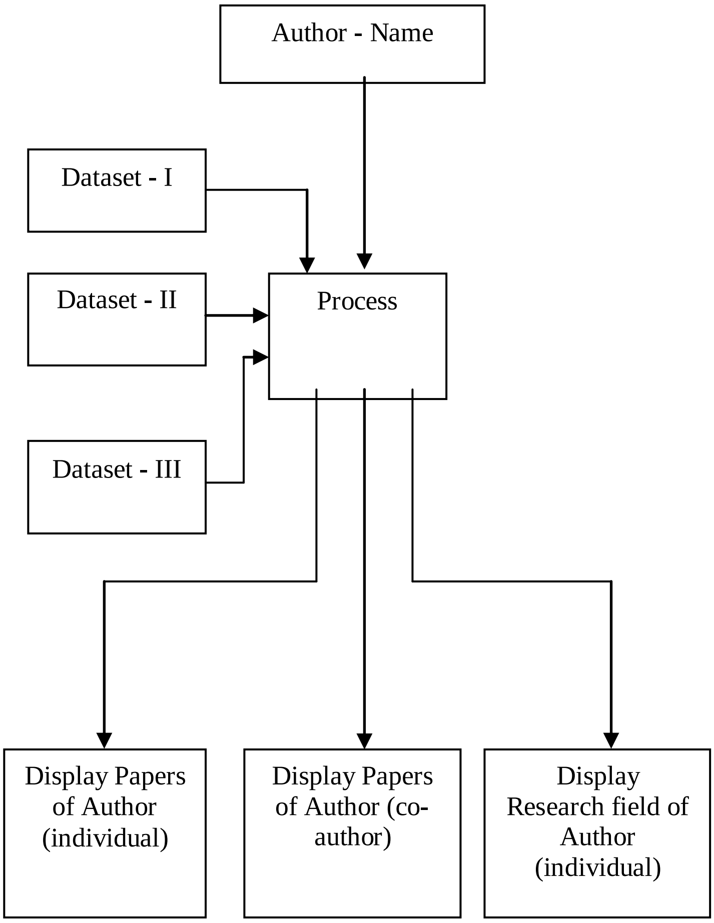

Author - Name
Dataset - I
Dataset - II
Process
Dataset - III
Display Papers of Author (individual)
Display Papers of Author (co-author)
Display Research field of Author (individual)

FIGURE 2-1 AUTHOR-TOPIC RELATION

In the above flow chart 3-1 the Author-name is the input of the interface, datasets includes:
Dataset – I: List of STOP WORDS
Dataset – II: List of PUNCTUATION MARKS
Dataset–III: Collection of PAPERS

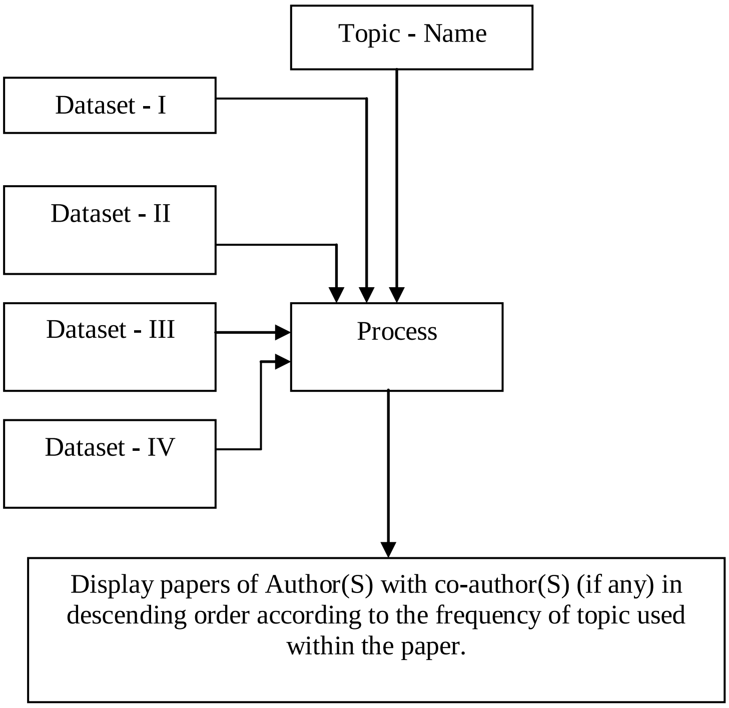

Topic - Name
Dataset - I
Dataset - II
Dataset - III
Process
Dataset - IV
Display papers of Author(S) with co-author(S) (if any) in descending order according to the frequency of topic used within the paper.

FIGURE 2-2 TOPIC-AUTHOR RELATION

In the above flow chart 3-2 the Topic-name is the input of the interface, datasets includes:
Dataset – I: List of STOP WORDS
Dataset – II: List of PUNCTUATION MARKS
Dataset–III: Collection of PAPERS
Dataset – IV: List of RELATED WORDS OF EACH TOPIC.

III. BASIC ALGORITHM

Text clustering is a technique for unsupervised document organization. Clustering methods are used to group documents into meaningful category. This project attempts to build a simple text clustering tool using the frequent term-based clustering algorithm. Text clustering is a useful method for navigating large sets of documents.

The basic algorithm used for the purpose of searching the author-topic relation there is the following basic algorithm for finding the frequency of repeated related works within the paper as shown below:

Zeroth Pass	Convert all input documents to UTF-8. This prevents encoding errors from creeping into the system during later stages.
First Pass	Lowercase all words and remove all punctuation. This gives us a simple stream of words which can be easily processed in the later stages. This stage does not impact the efficiency of the algorithms because we're using purely statistical methods for clustering. Sentence structure does not affect our algorithm.
Second Pass	Scan the input documents and remove all stop words. Stop words include common words in the English language. For example, words like "the", "a", "an" etc. appear in all documents. Including stop words in the clustering process impacts the accuracy of the clustering algorithm. A stop word list may be generated by taking the top n% words from the input

	data, or by downloading pre-generated stop word lists from the web. Also, remove all non-dictionary words from the input data. This prevents misspellings/slang from appearing in cluster names. Domain-specific dictionaries may be used for clustering domain-specific documents.
Third Pass	Create a mapping from the list of all words appearing in the document to a list of all documents that contain the words.
Fourth Pass	Here, the words appearing in top N% of the documents are taken as categories, and the rest of the words are discarded. Documents that do not contain any of the top N% of the words are filed under 'Uncategorized'.
Fifth Pass	This is the final pass. Here, sort all the documents categorized from the above pass, and arrange all the documents in the decreasing order according to the frequency of the given word within the document.

Content modeling Algorithm

The author-topic model described in this paper includes two types of searching methodology i.e. either mapping of author-name to find their papers either individually or with co-authors and to find his research field or mapping of topic-name to find all the papers related to that topic of all the authors starting from most famous author to least famous author.

a. ATP Mapping

Here we are given the name of an author, and we are interested in finding his research papers either individually or with any co-author and all interested in finding his research field as shown at the end in the figure 3-1.

b. TAP Mapping

Here we are given the topic name, and we are interested in finding all the papers on this topic of all the authors of this field in an order i.e. most popular author comes first while least popular author comes at last as shown at the end in the figure 3-2.

IV. EXPERIMENTAL RESULTS AND SETUP

To implement these two mapping models we had used the following tools:

FRONT END

Visual Studio 2005 with C#

BACK END

Here I had assumed some predefined datasets of my own but the system can be extended and uses the dataset of any digital library like CiteSeer.

DATASETS USED

A. Dataset for Stop words:

The dataset of stop words which are used in pass second and applied on all the paper's abstract, keywords, and conclusion for elimination is given below:

"a,able,about,across,after,all,almost,also,am,among,an,and,any,are,as,at,be,because,been,but,by,can,cannot,could,dear,did,do,does,either,else,ever,every,for,from,get,got,had,has,have,he,her,hers,him,his,how,however,i,if,in,into,is,it,its,just,least,let,like,likely,may,me,might,most,must,my,neither,no,nor,not,of,off,often,on,only,or,other,our,own,rather,said,say,says,she,should,since,so,some,than,that,the,their,them,then,there,these,they,this,tis,to,too,twas,us,wants,was,we,were,what,when,where,which,while,who,whom,why,will,with,would,yet,you,your"

B. Datasets for punctuation marks:

The dataset of punctuation marks which are used in pass first and applied on all the paper's abstract, keywords, and conclusion for elimination is given below:

',', '.', ':', ' ', '\n', '\0', '\r' ,'-','+','=','*','&','^','%','$','#','@','!','~','\"','\'','\\','1','2','3','4','5','6','7','8','9','0'

C. Dataset for research papers:

The dataset required in the mapping model is the set of papers in the predefined format as described above.

We had implemented the above mentioned two mapping models using the datasets and the interface used to extract the information to assist the new researchers is shown below:

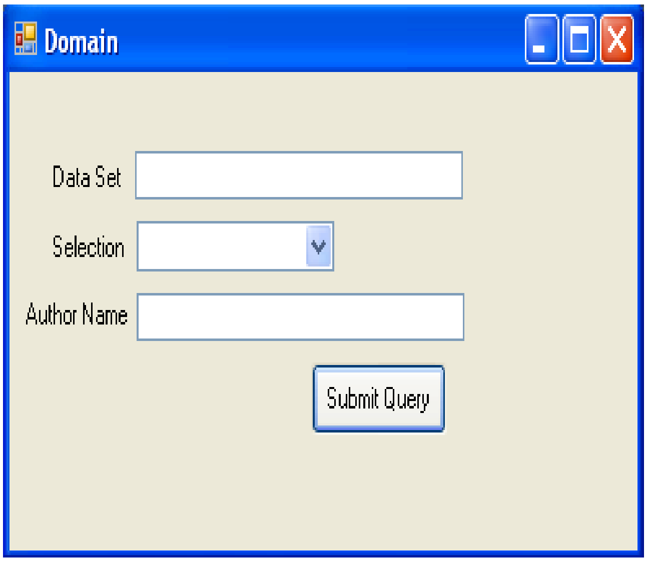
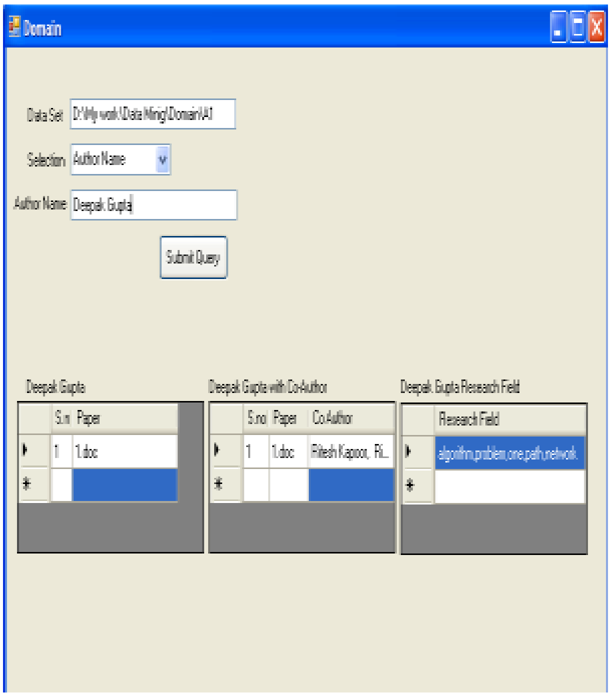


Figure 4-1 Front Interface

In the above interface there are two textboxes, one drop down box, and a submit button. The first textbox labelled Data Set is used to insert the complete path of the folder where the collection of papers of various research areas are stored, a drop down box labelled Selection is used to choose one of the following mapping model out of ATP and TAP and whatever model is selected the input is according entered in the third textbox labelled either Author Name or Topic and submit button is used to execute the query.

The interface for ATP model in which we have to extract information on the basis of author-name as shown below:



Data Set D:\My work\Data Minig\Domain\A1

Selection Author Name

Author Name Deepak Gupta

Deepak Gupta

S.n	Paper
1	1.doc

Deepak Gupta with Co-Author

S.no	Paper	Co Author
1	1.doc	Ritesh Kapoor, Ri...

Deepak Gupta Research Field

Research Field
algorithm,problem,one,path,network

Figure 4-2 Interface for ATP Model

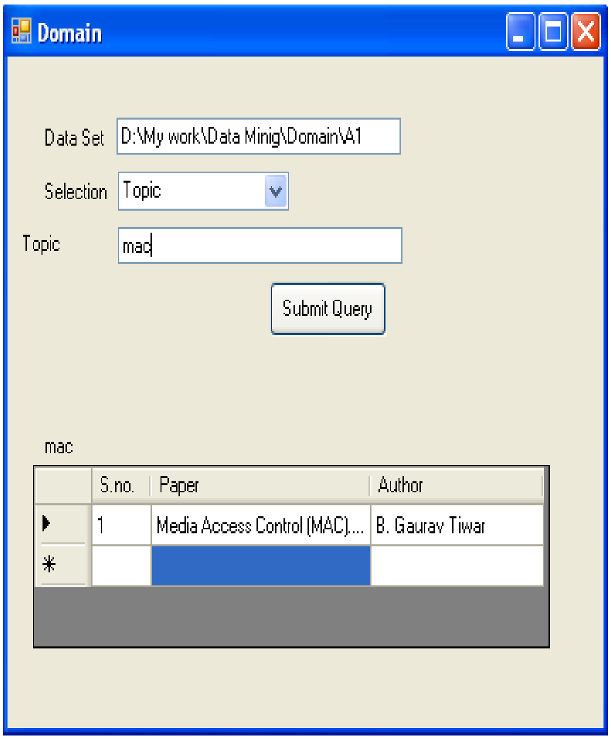
As discussed above the output of ATP mapping model is in the form of three tables as shown above in the screen shot, the three tables are used to give:

A. Generates all the papers of the author (either individual or with co-author(s)) with the paper's title.

B. Generates all the papers of the author only with the co-author(s) with the paper's title.

C. Generates the research field of the author and displays five most frequent words.

The interface for TAP model in which we have to extract information on the basis of topic-name as shown below:



Data Set D:\My work\Data Minig\Domain\A1

Selection Topic

Topic mac

mac

S.no.	Paper	Author
1	Media Access Control (MAC)....	B. Gaurav Tiwar

Figure 4-3 Interface for TAP Model

As discussed above the output of ATP mapping model is in the form of a table as shown above in the screen shot, the table is used to give:

A. Generates all the papers of all author according to the topic-name along with paper's title and author-name's (either individual or with co-author(s)).

V. CONCLUSION

The word data-mining is based on the metaphor in which nuggets of knowledge are sought inside a huge stack of unrelated information – the proposal being that the data-mining identify and refines something that is already present from the outset.

We have introduced a frequency based algorithm that can that can automatically extract information about authors, topics and gives relation about author, topic, document and words from large text corpora. The method uses a generative frequency Content model that links authors to observed words in documents. We demonstrated software which can be used to learn such content modelling paradigm from very large text corpora (including Abstract, Keywords & Conclusion) as a working example. We had also shown a case study of the probability based author-topic model which is applied successfully on the large text corpora. Content modelling paradigm was shown to extract substantial novel "hidden" information from the set of abstracts containing topic time-trends, author, document, topic, & words relation, and unusual papers for specific authors and so forth. Other potential applications not discussed here include recommending potential reviewers for a paper based on both the words in the paper and the names of the authors. Even though the underlying frequency based Content model is quite simple, and ignores several aspects of real-world document generation (such as topic correlation, author interaction, and so forth), it nonetheless provides a useful first step in understanding author-topic structure in large text corpora.

Future Work

The content model paradigm proposed and designed in this paper is only being executed on the predefined and assumed datasets as the real-time datasets are not accessible to the author. This proposed content model is easily pluggable and extendible in real-time datasets like CiteSeer digital library etc.

Our content model is ignoring a lot of useful facts that can potentially be advantageous toward the completion of this task. Polarity and sentiment of the reviews can provide some good clues for discovering the aspects. There may also be some benefits to use some improved initial sets of aspect by other types of clustering such as Spectral clustering or K-Means both jointly or independently with SS-LDA. Whereas literature-cantered networks are developed to ask questions about publication performance, a dissimilar (and simpler) type of network is more suited for asking questions about collaboration performance: every researcher I is a node; if Ii and Ij have jointly co-authored one paper or article, they are joined by a non-directed link of strength a or 1. If they have co-authored two articles, the link has strength b or 2, and so on. Again, a very large number of characteristics can be linked with each researcher/node: internal features, inherent network characteristics, and external information. One can even utilize a content that is obtained from the researcher -literature networks, e.g., if researcher Ii stands in n-th degree relation to another researcher Ij in an researcher -literature network, then this information can be used as one of the characteristics in the collaboration system. The study of scientific collaboration is a complete field in itself (Sonnenwald, 2007), and there are a lot of dissimilar ways in which collaboration networks can be analyzed. One can attempt to understand which factors decide whether two individuals will collaborate together (resulting in a joint publication). One can also inspect networks as they grow over time. These basic modelling studies set the stage for creating user-friendly tools that will allow an individual to search potentially high-quality collaborators for a specified problem. Since one individual might be an excellent potential collaborator for a huge number of individuals, far too many to work with all at once, it is essential to consider constraints and limiting factors as well.

Author name disambiguation has strategic significance that goes far beyond knowing who-wrote-what. The type of collaboration networks is merely the simplest example of how disambiguation information can underlie the development of new resources and tools that open up entirely dissimilar type of researcher. As library and information science becomes more and more person-centred, and not just document-centred, we will be expecting to see ripples that will affect the semantic web, world of publishing, the indexing of data collections, and the design of search engines.

REFERENCES

[1] T. F. Lunt, J. van Home, and L. Halme. Analysis of computer system audit trailsinitial data analysis. Technical Report TR-85009, Sytek, Mountain View, California,September 1985.

[2] J. van Horne and L. Halme. Analysis of computer system audit trails final report.Technical Report TR-85007, Sytek, Mountain View, California, May 1986.

[3] Peter G. Neumann. Security and integrity controls for federal, state, and local computersaccessing NCIC. Technical report, SRI International, 333 Ravenswood Avenue, MenloPark, CA 94025, June 1990.

[4] Alfonso Valdes and Debra Anderson. Statistical methods for computer usage anomalydetection using NIDES. In Conference on Rough Sets and Soft Computing, November1994.

[5] Boyd-Graber, J. & Blei, D., 2009. Syntactic Topic Models. In Neural Information ProcessingSystems.

[6] Branavan, S., Chen, H., Eisenstein, J. & Barzilay, R., 2008. Learning Document-Level Semantic Properties from Free-text Annotations. In Proceedings of ACL.

[7] Lin, J., 1991. Divergence measures based on the Shannon entropy. In IEEE Transactions onInformation Theory.

[8] Mann, G. & McCallum, A., 2008. Generalized Expectation Criteria for Semi-Supervised Learning of Conditional Random Fields. In ACL.

[9] Mccallum, A., Corrada-Emmanuel, Andres & Wang, X., 2005. Topic and Role Discovery in Social Networks. In Proceeding of IJCAI.

[10] Blei, D.M. & McAuliffe, J., 2007. Supervised topic models. In Advanced In NIPS.

Blei, D.M., Ng, A.Y. & Jordan, M.I., 2003. Latent Dirichlet Allocation. In Journal of Machine Learning Research.

[11] Minka, T. & Lafferty, J., 2002. Expectation-propagation for the generative aspect model. In Proceedings of UAI.

[12] Newman, D., Chemudugunta, C. & Smyth, P., 2006. Statistical entity-topic models. In: 10th ACM SigKDD conference knowledge discovery and data mining (Seattle, 2004)

[13] Mark Steyvers, Padhrai Smyth, Thomas Grihffiths, Probabilistic Author-Topic Models for Information Discovery.

[14] H.S. Javitz and A. Valdes. The SRI statistical anomaly detector. In Proceedings of the1991 IEEE Symposium on Research in Security and Privacy, May 1991.

[15] J. P. Anderson. Computer security threat monitoring and surveillance. Technical report, James P. Anderson Company, Fort Washington, Pennsylvania, April 1980.

[16] T. F. Lunt, J. van Horne, and L. Halme. Automated analysis of computer system audit trails. In Proceedings of the Ninth DOE Computer Security Group Conference, May1986

[17] Waterman, D.A, (1984) A guide to Expert Systems, Reading, Addison-Wesly, Massachusetts.

[18] Blei, D.M. & Jordan, M.I., 2003. Modeling annotated data. In SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.

[19] Chang, J. & Blei, D., 2009. Relational Topic Models for Document Networks. In Artificial Intelligence and Statistics.

[20] Cohen, J., 1960. A coefficient of agreement for nominal scales. In Education andPsychological Measuremen.

[21] Deerwester, S. et al., 1990. Indexing by latent semantic analysis. In Journal of the AmericanSociety for Information Science.

[22] Goldwater, S., Griffiths, T.L. & Johnson, M., 2006. Contextual Dependencies in Unsupervised Word Segmentation. In Proceedings of Coling/ACL.

[23] Griffiths, T.L. & Steyvers, M., 2004. Finding scientific topics. In Proc Natl Acad Sci U S A. Griffiths, T.L., Steyvers, M., Blei, D.M. & Tenenbaum, J.B., 2005. [24] Integrating topics and Syntax. In Advances in NIPS 17.

[25] Gruber, A., Rosen-Zvi, M. & Weiss, Y., 2007. Hidden Topic Markov Models. In ArtificialIntelligence and Statistics.

[26] Haghighi, A. & Klein, D., 2007. Unsupervised Coreference Resolution in a Nonparametric Bayesian Model. In Association for Computational Linguistics.

[27] Hofmann, T., 1999. Probabilistic latent semantic analysis. In Proc. of Uncertainty in Artificial Intelligence, UAI’99.

[28] Hu, M. & Liu., B., 2004. Mining and summarizing customer reviews. In Proceedings of SIGKDD.

[29] Levin, E. & Sharifi, M., 2006. Evaluation of Utility of LSA for Word Sense Discrimination. In Proceedings of HLT/NAACL.

[30] Blei, D. & Lafferty, J., 2006. Dynamic topic models. In Proceedings of the 23rdInternational Conference on Machine Learning.

[31] Blei, D. & Lafferty, J., 2007. A correlated topic model of Science. In Annals of AppliedStatistics.

[32] Teresa Lunt. Detecting intruders in computer systems. In 1993 Conference on Auditingand Computer Technology, 1993.

[33] Next-generation Intrusion Detection Expert System by Debra AndersonThane Frivold Alfonso Valdes Computer Science Laboratory 1995.

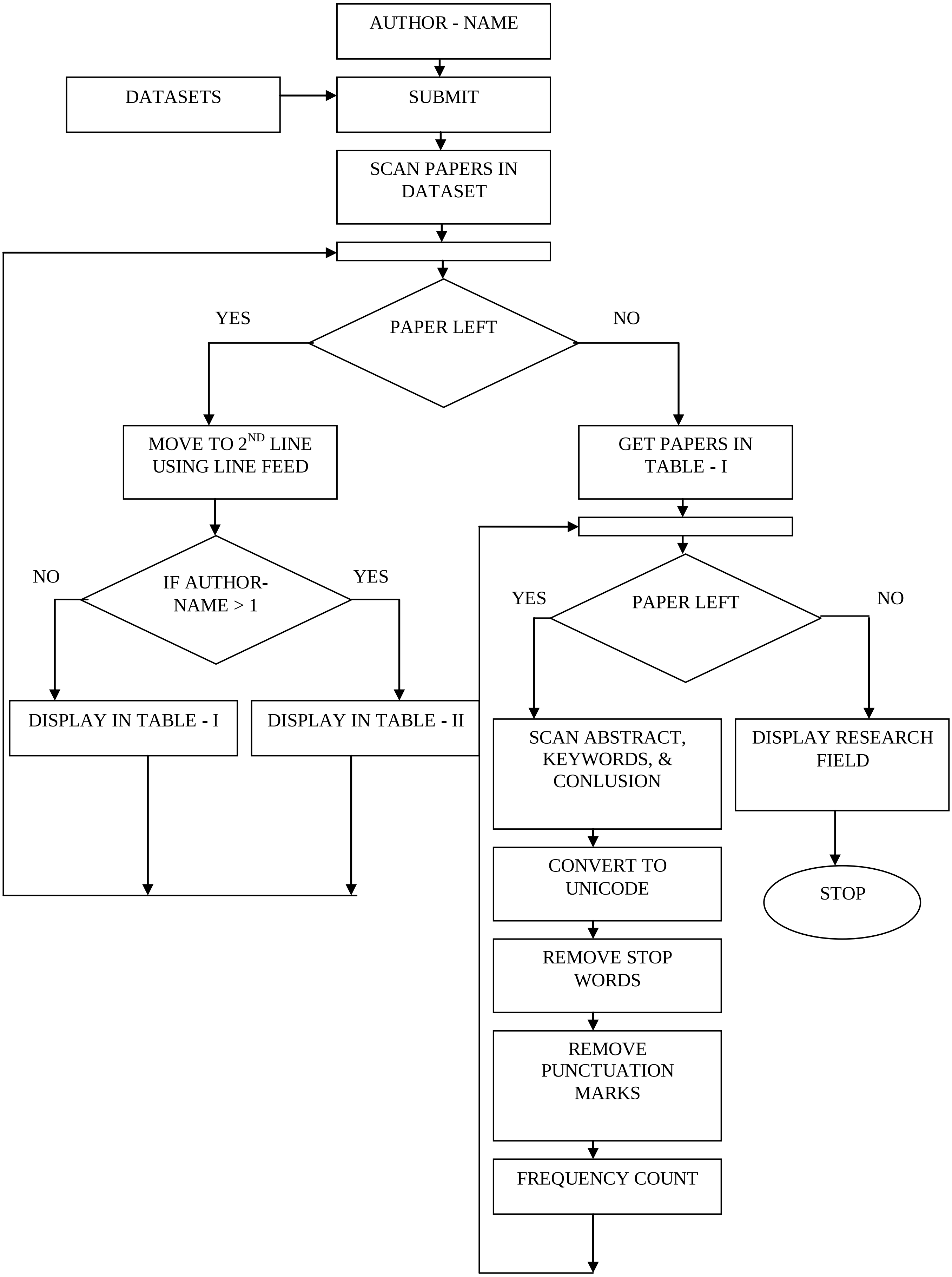


Figure 3-1 ATP Model

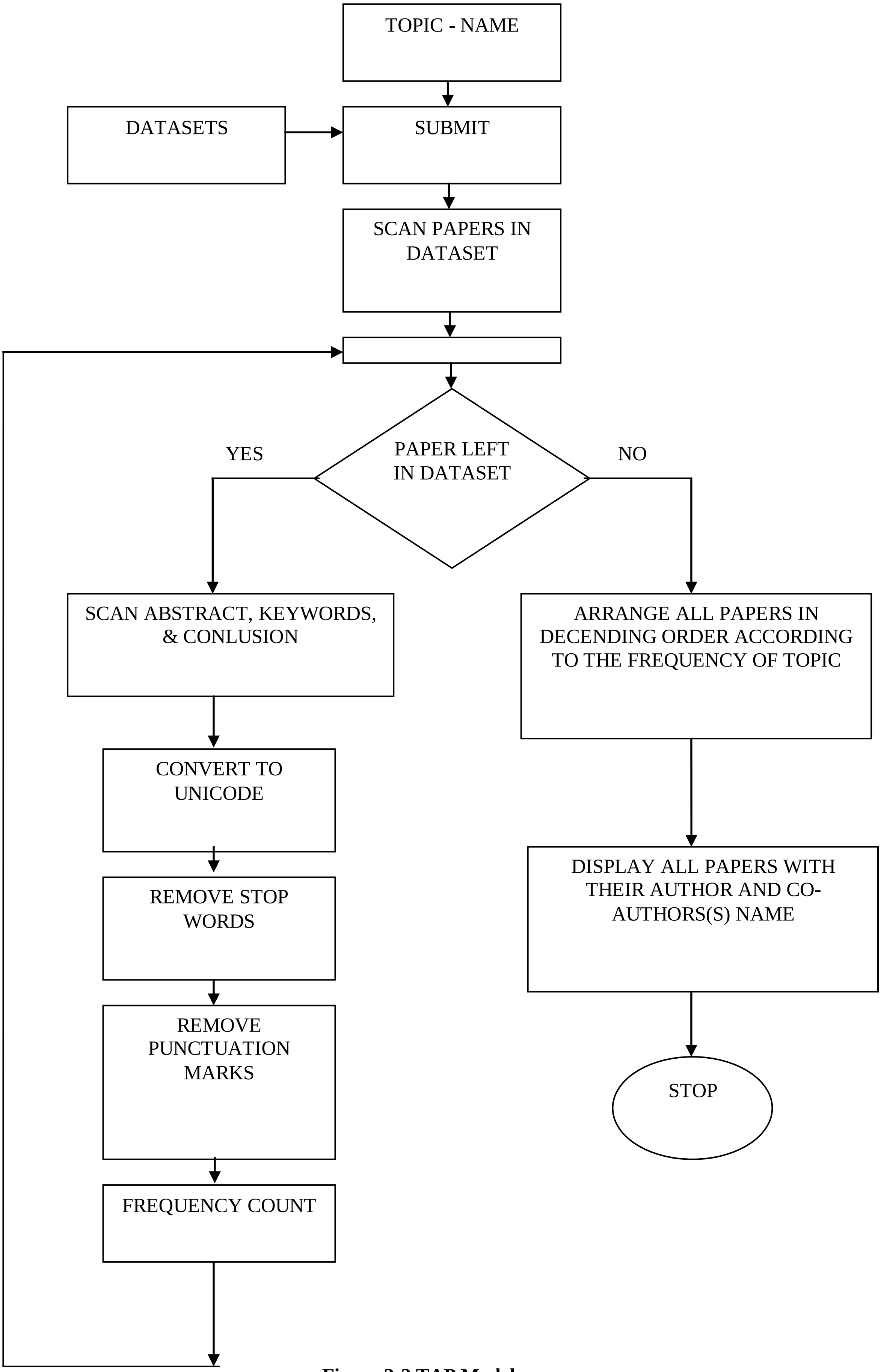

TOPIC - NAME
DATASETS
SUBMIT
SCAN PAPERS IN DATASET
PAPER LEFT IN DATASET
YES
NO
SCAN ABSTRACT, KEYWORDS, & CONLUSION
CONVERT TO UNICODE
REMOVE STOP WORDS
REMOVE PUNCTUATION MARKS
FREQUENCY COUNT
ARRANGE ALL PAPERS IN DECENDING ORDER ACCORDING TO THE FREQUENCY OF TOPIC
DISPLAY ALL PAPERS WITH THEIR AUTHOR AND CO-AUTHORS(S) NAME
STOP

Figure 3-2 TAP Model