DATA MINING DESIGNING WITH LARGE DATA SET

[1]Prageet Bajpai
Computer Science & Engineering
SSCET Bhilai
Chhattisgarh India

[2]Siddharth Choubey
Computer Science & Engineering
SSCET Bhilai
Chhattisgarh India

[3]Abha Choubey
Computer Science & Engineering
SSCET, Bhilai
Chhattisgarh, India

Abstract— — Data warehouse is an important contemporary issue for many organizations and is relatively a new field in the realm of information technology. As data warehousing is a new field, little research has been done regarding the characteristics of academic data and the complexity of analyzing such data. Educational institutions measure success very differently from business-oriented organizations and the analyses that are meaningful in such environments pose unique problems in data warehousing. The purpose of this thesis is to provide a security of data ware house In the present work we have introduced a vernam cipher bit wise encryption method.

INTRODUCTION

Text mining is a practice that is utilized to find advantageous in arrange mention from large amount of data sets. Data mining has guidelines known as frequent pattern and association rule that is essential for finding frequent patterns. Text Mining is the recognition by computer of new, previously unidentified in arrange mention, by automatically mining in arrange mention from different written resources. Text mining techniques are the fundamental and permitting tools for efficient organization, triangulation, retrieval and summarization of large file quantity. With more and more text, in arrange mention are distribution around on Internet, text mining is rising in importance. Text clustering and text classification are two important tasks in the field of text mining.

Text Clustering is to find out the groups in arrange mention from the text file and cluster these file into the most relevant groups. Text clustering clusters the file in an unsupervised way and there is no label or class in arrange mention. Clustering techniques have to determine the connections between the file and then based on these connections the files are bunched. Given a enormous volumes of files, a superior document clustering techniques may organize those huge statistics of documents into meaningful groups, which permit further browsing and navigation of this quantity be much easier (B.Liu, M.Hu and J. Cheng, 2005). A basic idea of text clustering is to find out which types of documents have many words in common and place these types of documents with the most words in mutual into similar group.

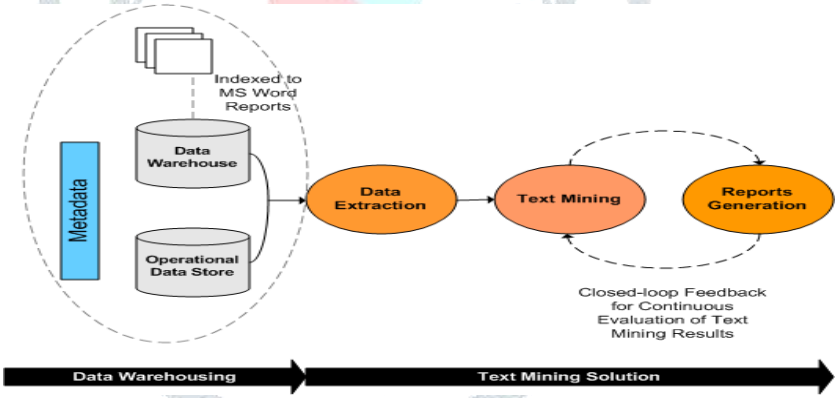

Figure 1 Process of Text Mining Block Diagram

Text Classification is to establish the files into predefined classes with meaningful labels. As text classification wants the facts about those predefined categories, it is applied in a supervised way.

Eighty percent of the inarrangemention in the world is currently stored in amorphous textual arrangement. Although method such as Natural Language Processing (NLP) can complete restricted text analysis, there are currently no computer programs available to investigate and interpret text for the different inarrangemention extraction wants. Thus text mining is a dynamic and unindustrialized region. The world is fast becoming inarrangemention comprehensive, in which specialized inarrangemention is being poised into extremely large data sets. For instance, Internet contains a large amount of online text files, which rapidly change and grow. It is nearly dreadful to manually organize such vast and quickly evolving inarrangemention. The requisite to extract useful and relevant inarrangemention from such bulky data sets has led to a significant requirement to develop computationally competent text mining algorithms (A.M.Popescu and O. Etzioni , 2005). An instance, problem is to automatically dispense natural language text files to predefined sets of categories grounded on their contented. Other instances of problems involving large data sets comprise searching for targeted inarrangemention from technical citation databases (e.g. MEDLINE); search, filter and classify web pages by topic and routing relevant email to the proper addresses.

Text mining is the involuntary and semi-automatic removal of implicit, previously indefinite, and hypothetically useful inarrangemention and patterns, from a large amount of amorphous textual data, such as natural-language text (D. Kerr, H. Mousavi, and M. Iseli ,2013). In text mining, every file is represented as a vector, whose dimension is almost the number of diverse keywords in it, which can be very large. One of the major contests in text mining is to categorise textual data with such superior dimensionality. In adding up to high dimensionality, text-mining algorithms would also deal with word ambiguities such as pronouns, synonyms, and deafening data, spelling mistakes, abbreviations, acronyms and inadequately structured text. Text mining algorithms are of two types: Supervised learning and unsupervised learning. Support vector machines (SVMs) are a set of supervised learning approaches utilized for classification and reversion. Nonnegative matrix factorization is an unsupervised learning method.

1.1.1 Supervised Learning

Supervised learning is a technique in which the algorithm customs predictor and target attribute value couples to learn the predictor and target value relation. Support vector machine is a supervised learning process for making a decision function with an exercise dataset. The training data contain pair of predictor and target values. Each predictor value is considered with a target value. If the algorithm can imagine a categorical value for a target attribute, it is known as classification task. Class is an instance of a definite variable. Positive and negative be capable to be two values of the inflexible variable class. Categorical values do not have partial ordering. If the algorithm can imagine a numerical value then it is called regression. Numerical values have limited ordering.

1.1.2 Unsupervised Learning

Unsupervised learning is a technique in which the algorithm practices only the predictor attribute values. There are no object attribute values and the learning chore is to add some understanding of pertinent structure patterns in the data. Each row in a data set signifies a point in n-dimensional space and unsupervised learning algorithms examines the relationship among these numerous points in n-dimensional space. Instance of unsupervised learning are clustering, density approximation and characteristic extraction.

1.1.3 Learning Machine

Support vector machine is a type of learning machine. A learning machine is given a training set of sample or inputs with connected labels or output values. Generally the example is in the form of characteristic vectors, so that input is a subset of Rn. For example, consider an input X = (x1, x2, ... xn), in which X belongs to an n-dimensional vector space Rn and x1, x2, … xn are the component of the vector X. X is totally assigned to the positive class, if f(X) greater than or equal to 0, and to the negative class if f(X) less than 0. In this case, the function f(X) is a decision function. Each vector has the target attribute of $Y \in \{-1, +1\}$, where i = 1 to n. and -1 and +1 are negative and positive classes respectively. A learning machine learns the mapping $X \Rightarrow Y$, which can be characterised by a set of probable mappings, $X \Rightarrow f(X,\alpha)$, where α is a set of strictures for the function f(X). For a given input of X and a choice of α, the machine will constantly give the similar output. Since there are solitary two classes, the objective here is to build a binary classifier from the training samples (predictor-target value pairs for learning the machine), which has a small possibility of misclassifying a testing sample (predictor-target value pairs for challenging the machine). For the file cataloguing difficulty, X is a characteristic vector for a file. This characteristic vector comprises of frequencies of diverse keywords and Y is the user-defined category.

II. PREVIOUS WORK

Many data mining techniques have been proposed for mining useful patterns in text documents. How to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis.

Zhong N. et al (2012) presented an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered the patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics
demonstrated that the proposed solution achieves encouraging performance.

Luepol Pipan maekaporn (2013), presented a novel pattern mining approach to RF. This approach mined patterns in both positive and negative feedback and then classified them into clusters to find user-specific patterns. They also proposed a novel pattern deploying method that effectively used the discovered patterns for improving the performance of searching relevant documents. Experiments are conducted on Reuters Corpus Volume 1 data collection (RCV1) and TREC filtering topics. The results shown that the proposed approach achieves promising performance comparing with state-of-the art term-based methods and pattern-based ones.

They also applied a novel pattern deploying the strategy to improve the performance of frequent patterns in text. They evaluated the proposed approach by using it to discover high-quality features in relevance feedback for improving the information filtering. Their results on RCV1 data collection and TREC filtering topics confirmed that the best improvements are obtained by our approach compared to state-of-the-art term-based methods and pattern-based ones.

Bhushan Inje, Ujawla Patil (2014) examined and investigated this fact with considering several states of art data mining methods that gives satisfactory results to improve the effectiveness of the pattern. Here they implemented the pattern detection method to solve problem of term-based methods and improved result which is helpful in information retrieval systems. Their proposal was also evaluated for several they'll distinguish domain, offering in all cases, reliable taxonomies considering precision and recall along with F-measure. For the experiment, they used Reuters (RCV1) dataset and the results show that they improved the discovering pattern as compared to previous text mining methods. The results of the experiment setup show that the keyword-based methods not give better performance than pattern-based method. The results also indicated the removal of meaningless patterns not only reduces the cost of computation but also improved the effectiveness of the system.

In this paper, they have investigated the existing data mining methods with respect to the alternating approach for finding relevant pattern in large documents collection; some research work have been used phrases rather than individual words. However, the effectiveness of the text mining systems was not improved very much. The likely reason is that, a phrase-based method has "lotheyr consistency of assignment and lotheyr document frequency for terms". Hence, in this paper, they presented a concept for mining text documents for sequential patterns. Instead of using single words, they used pattern-based taxonomy (is-a) relation to represent documents. By pruning meaningless (negative) patterns, which have been proven the source of the 'noise' in this study, the problem of over fitting is solved and the experimental results, which shown the encouraging outcomes, are achieved. The results of the experiment show that the keyword-based methods not gives better performance compare to pattern-based method. The results also indicated that removal of meaningless patterns not only reduced the cost of computation but also improves the effectiveness of the system.

III. PROBLEM IDENTIFICATION

Large file corpus may manage to pay for a lot of useful information to people. But it is also a challenge to come through out the beneficial sequence from huge number of documents. Especially with the detonate of knowledge around the cyber-world, company and establishments demand efficient and ineffectual ways to systematize the large manuscript corpus and make later directing and browsing to be converted into more easy, friendly and efficient. An obvious distinctive of large file corpus is the enormous sizes of forms. It is almost impossible for a man

to read from side to side all the papers and come across out the qualified for a definite topic. How to organize large document corpus is the problem we concern.

It was attainable that the formal classification of the problem of sequential pattern mining and its use to demining the web log. Known (i) a set of consecutive records (called sequences) on behalf of a successive file D; (ii) a least amount provision edge called min sup ξ ; and (iii) a set of k only one of its type of objects or events I = {i1, i2, . . . , ik}. The difficulty of mining consecutive patterns is that of declaration the set of all recurrent progressions Sin to the fixed order record D of items I at the given min sup.

In many real applications more than ever in compact data with long recurrent patterns enumerating all potential2L – 2 subsets of an L length pattern is infeasible. A sensible solution is recognizing a slighter emissary set of patterns from which all other frequent patterns can be subsequent. Maximal recurrent patterns (MFP) form the lowest representative setof patterns to engender all frequent designs. In honorable, the MFP are those forms that are frequent but not an iota of their supersets are frequent. The problem of highest repeated forms mining is finding all MFP in D with high opinion to σ.

The difficulty of repeated designs mining from a large amount of data is making a huge number of patterns disappointing the smallest amount sustain verge, specifically when min_sup σ is detailed low. This is for, all sub-pattern of a recurrent pattern are recurrent as well. Subsequently a long design contains a number of shorter frequent sub patterns. Assorted category of frequent designs can be excavation from various kinds of data sets. In this make inquiries, we utilize item sets (sets of items) as a files set and the wished-for technique is for frequent item set mining, that is, the mining of common from transactional records sets. However, it can be comprehensive for extra kinds of frequent patterns.

The difficulty is more frequent than not decayed into two sub difficulties.

1. One is to find those item sets whose occurrence goes outside a predefined inception in the database; those item sets are describe frequenter big item sets.
2. The second difficulty is to yield participation rules from those huge item sets with the restriction of minimal self-confidence.

The most noteworthy challenging topic in text mining occur from the difficulty of a usual language itself. The natural language is not free from the obscurity problem. Obscurity means the wherewithal of being understood in two or more achievable senses or ways. Ambiguity gives a natural language its flexibility and usability, and consequently, therefore it cannot be completely removed from the natural language. One word may have multiple meanings. One phrase or sentence can be understood in various ways, thus various connotations can be obtained. Although a number of researches have been demeanour in resolving the ambiguity problem, the work is still undeveloped and the planned approach has been dedicated for a detailed field. On the other hand, maximum of the IE systems that involve semantic investigation exploit the simplest part of the whole spectrum of domain and task information, that is to say, named entities.

According to authors, IE does a more limited task than full text thoughtful. He pointed that in full text understanding, all the information in the text is accessible, whereas in information pulling out, the semantic range of the output, the kindred will be existing are delimited. However, the on the increase need for IE application to domains such as well-designed genomics requires extra text understanding. Named entity recognition (NER) describes an classification of entities in free text. For example, in biomedical field, objects would be gene, protein names and drugs. NER often forms the opening point in a text mining system, association that when the correct entities are predictable, the search for patterns and relations between entities can begin.

Large article corpus may afford a lot of constructive information to people. But it is also a confront to find out the helpful information from huge number of documents. Particularly with the explode of knowledge in the region of the cyber-world, corporates and administrations demand well-organized and efficient ways to organize the bulky document corpus and build later navigating and browsing become more trouble-free, friendly and competent.

An obvious attribute of large document quantity is the huge volumes of documents. It is almost not possible for a man to read throughout all the documents and find out the comparative for a specific topic. How to systematize large document corpus is the predicament we concern.

IV. METHODOLOGY

The methodology which has been proposed for the solution of the problems identified in the project is as shown in the Figure 4.1

To perform the experiment, we will need the MATLAB Tool for Simulation.

Methodology in Detail

Algorithm : Training the system to identify which class the text document belongs.	
Step- 1	The system is trained in three classes viz Historical Class, Constitutional Class and Geographical Class.
Step- 2	The system has to be trained on the basis of the documents which are related to these predefined classes.
Step- 3	The system is trained by extracting several keywords from the document related to a particular field then the keywords that are unique with respect to each are stored to classify the given text document.
Step – 4	Then after the text document is uploaded into GUI to find the class to which the text document belongs.

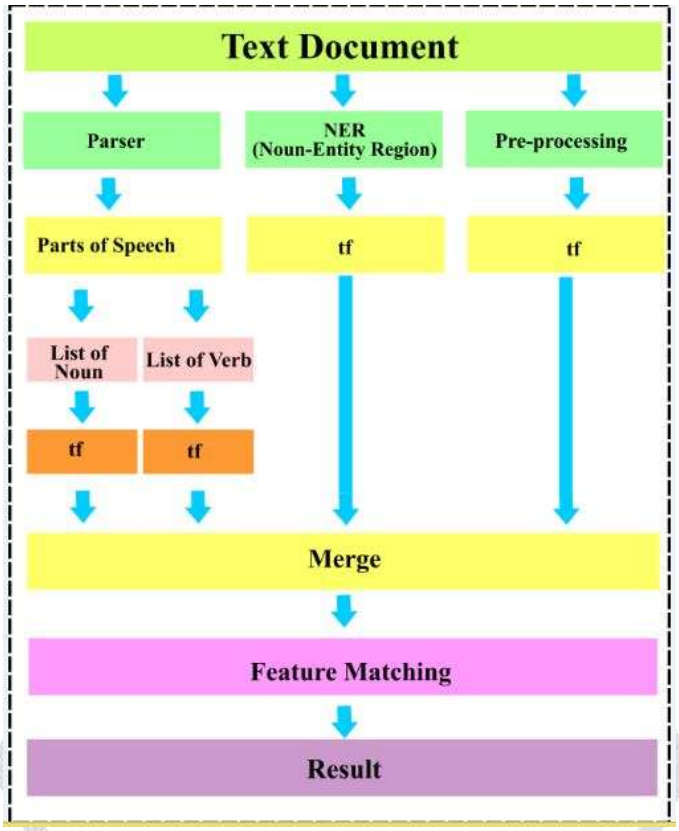

Figure 2 Flowchart of the Methodology

Algorithm 4.2: Training Parse the text document using Stanford parser

Step- 1	Parsing or syntactic analysis is the process of analyzing a string of symbols, either in natural language or in computer language, conforming to the rules of a formal grammar.
Step- 2	The parsing is done by Stanford parser, which converts the text document into parse tree.
Step- 3	One way to parse the text document is to parse each sentences of the text document individually and saving the output but the drawback is that it involves loading the Stanford parser several times.
Step – 4	The headwords are saved for feature matching at later stage. Each words in the PTs is assigned with an ID to make system able of uniquely addressing words in the text document. This is required to avoid confusion amongst repeated words.
Step – 5	The headwords are saved for feature matching at later stage. Each words in the PTs is assigned with an ID to make system able of uniquely addressing words in the text document. This is required to avoid confusion amongst repeated words.
Step – 6	The PT contains list of noun and verb which will be used in feature matching and by that concluding which class the given text document belongs.

Algorithm 4.3: Apply Noun Entity Region (NER).

Step- 1	Noun entity region (NER) (also known as entity identification, entity chunking and entity extraction) is a sub-task of information extraction that seeks to locate and classify elements in text into pre-defined category such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.
Step- 2	To evaluate the quality of a NER system's output, several measures have been defined. While accuracy on the token level is one possibilities, it suffers from two problems: the vast majority of tokens in real-world text are not part of entity names as generally defined,

	so the baseline accuracy (always predict "not an entity ") is extravagantly high, typically >90%; and miss predicting the full span of an entity name is not properly penalized

Algorithm 4.4: Pre – processing of the text document.

Step- 1	Data may or may not have quality problems that need to be addressed before applying a data mining techniques.
Step- 2	The Data may be irrelevant or duplicate, thus pre-processing is necessary.
Step- 3	Pre-processing may be needed to make the text more suitable for text mining. There are a number of different tools and methods used for pre-processing.
Step – 4	Text preprocesses is the initial step of text mining which reads one text document at time and process it. This step divides into following major three subtasks- **Tokenization** Generally text document contain multiples sentences. So this process divides whole sentence into words by removing comma, spaces, punctuations etc. **Stop Word Removing** This process removes stop words such as “the”, “are”, “a” or any tags like HTML tag etc. **Stemming** Stemming is applied after stop word removal by reducing the word to its root word. E.g. “playing”, “played” are stemmed to “play”.
Step – 5	After pre-processing of the text document several headwords are generated. These headwords are saved to categorize the text document to the class it relates.

Algorithm 4.5: Separate the list of noun and the list of verbs from the document.

Step- 1	Separate the list of nouns and the list of verbs generated through that parser.
Step- 2	There are lot of ‘ing’ and ‘es’ or‘s’ like words which have to be separated from the word to make it a proper word.
Step- 3	There is a facility of word net , which recognizes the synonyms of the words and considers into the same word count, which is a very significant step in the method.
Step – 4	Out of these, the nouns and the verbs are separated and made a parse tree.
Step – 5	The list of noun is also generated by NER (Noun-Entity Region) which is termed as Main Part – I

Algorithm 4.6: Merge all the features

Step- 1	After separating all the headwords viz noun and verb words generated by parser, noun words generated from NER (Noun-Entity Region) and the word generated after pre-processing, merge all the headwords extracted by system in order to identify that the given document belongs to which pre-defined class viz Historical Class, Constitutional Class and Geographical Class.

Algorithm 4.7: Match the features with the unique headwords of the predefined classes viz Historical Class

Step- 1	The Features are matched with the trained system, which matches the noun and verb present in the text document with that of the list of unique headwords of the classes which are pre-defined and tells us that in which class the text document belongs.

Step- 2 If the text document doesn't match with any of the classes, then it gives the result, no matching found.

V. CONCLUSION

From the Research work, we came into a conclusion that mining the semantic information from free text document provides the enabling technology for a host to identify the class to which the text document belongs. The NER (Noun Entity Region) has been used to identify the noun keywords using classifier uniquely viz 3 – Class classifier, 4 – class classifier and 7 – class classifier. By using the concept of Parsing and NER, the text document has been classified to the predefined class to which the given text document belongs by merging the MP – I (List of noun) and MP – II (List of verb) and matched it with the stored headwords to conclude which class the document belongs to. The use of parser to convert the text document to parse tree increased the accuracy of the work.

V. FUTURE SCOPE

The Work is carried out with MATLAB as a Simulation tool, maybe there are some possibilities of increasing of accuracy rate if it is carried out by NS – 2 Simulator. As a future work, this solution can be enhanced by training the system in a broader way containing large number of documents and Classifiers. The application of Wordnet is also limited here, consisting of limited synonyms of each word, which can be enhanced to observe more accuracy.

IV. REFERENCES

[1] H. Mousavi, D. Kerr, M. Iseli, and C. Zaniolo. Harvesting domain specific ontologies from text. ICSC, 2014.

[2] H. Mousavi, S. Gao, and C. Zaniolo. Discovering attribute and entity synonyms for knowledge integration and semantic web search. SSW, 2013.

[3] H. Mousavi, S. Gao, and C. Zaniolo. Ibminer: A text mining tool for constructing and populating infobox databases and knowledge bases. PVLDB, 6(12):1330–1333, 2013.

[4] H. Mousavi, D. Kerr, M. Iseli, and C. Zaniolo. Deducing infoboxes from unstructured text in wikipedia pages. In CSD Technical Report #130001, UCLA, 2013.

[5] M. Atzori and C. Zaniolo. Swipe: searching wikipedia by example. In WWW, pages 309–312, 2012.

[6] T. Lee, Z. Wang, H. Wang, and S. won Hwang. Web scale taxonomy cleansing. PVLDB, 4(12):1295–1306, 2011.

[7] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, Jurafsky, and C. D. Manning. A multi-pass sieve for coreference resolution. In EMNLP, 2010

[8] R. Hahn, C. Bizer, C. Sahnwaldt, C. Herta, S. Robinson, M. Burgle,¨ H.Duwiger,¨ and U. Scheel. Faceted wikipedia search. In BIS, 2010..

[9] J.Han et al., Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery, 8, 53–87, 2004.

[10] Kluwer. M.Hu, and B.Liu, Mining and Summarizing Customer Reviews, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), USA, 2004, pp. 168 – 177.

[11] B.Pang, L.Lee, and S.Vaithyanathan, Thumbs up? Sentiment Classification Using Machine Learning Techniques, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02), USA, 2002, pp. 79 –86.

[12] B.Liu, M.Hu and J. Cheng, Opinion Observer - Analyzing and comparing opinions on the Web, in: Proceedings of the 14th International Conference on World Wide Web (WWW'05), Japan, 2005, pp. 342-351.

[13] A.M.Popescu and O. Etzioni, Extracting Product Features and Opinions from Reviews, Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP'05), Canada, 2005, pp. 339 – 346.

[14] X.Ding, B. Liu and S.Y.Philip, A Holistic Lexicon-Based Approach to Opinion Mining, in: Proceedings of the first ACM International Conference on Web search and Data Mining (WSDM'08), California, USA, 2008, pp. 231-240.

[15] M. Abulaish, Jahiruddin, M. N. Doja and T. Ahmad, "Feature and Opinion Mining for Customer Review Summarization", PReMI 2009, Lecture Notes in Computer Science, vol. 5909, pp. 219–224, Springer-Verlag Berlin Heidelberg 2009.

[16] M. Atzori and C. Zaniolo. Swipe: searching wikipedia by example. In WWW, pages 309–312, 2012.

[17] B.Pang, and L. Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, in: Proceedings of ACL 2004, 2004, pp. 271-278.

[18] R. Alur and P. Madhusudan. Adding nesting structure to words. In Developments in Language Theory, 2006.

[19] M. Atzori and C. Zaniolo. Swipe: searching wikipedia by example. In WWW, pages 309–312, 2012.

[20] E. Charniak and M. Elsner. Em works for pronoun anaphora resolution. In EACL, pages 148–156, 2009.

[21] S. A. Cook. The complexity of theorem-proving procedures. In STOC, pages 151–158, 1971.

[22] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. Gate: an architecture for development of robust hlt applications. In Recent Advanced in Language Processing, pages 168–175, 2002.

[23] B. J. Grosz, S. Weinstein, and A. K. Joshi. Centering: a framework for modeling the local coherence of discourse. Comput. Linguist., 21(2):203–225, June 1995.