ISSN:0975-9646

Research Challenges in Extracting Personal Name Pseudonyms from the Web

Sivagaminathan Ganesan

Senior Faculty, IBRI College of Technology
Department of Information Technology
Directorate of Technological Education
Ministry of Manpower
Sultanate of Oman

***Abstract*— A person may have multiple personal name aliases on the web. Identifying aliases of a name is useful in information retrieval and knowledge management, sentiment analysis, relation extraction and name disambiguation. The objective of detecting aliases from the web is to retrieve all the information pertaining to a personal name whose content is described with different nick names in different documents of web. As of now, web contains aliases of popular personalities in various domains like sports, politics, medicine, music, cinema etc., and does not contain alias information about common man. Recently, there are proven methods of extracting aliases through lexical pattern based retrieval tested using real-world name-alias pairs in Japanese and English as training data related to limited domains. In this paper, we discuss on how this information retrieval process has been grown now, and what the future directions are and also the scope of alias extraction in Inter-disciplinary fields.**

Keywords: Information Retrieval, Mnemonic name, name disambiguation, Word Sense Disambiguation, Lexico-syntactic pattern , Natural Language Processing, Semantic Similarity

1. INTRODUCTION

Finding information about people in the web is one of the day-to-day activities among Internet users. Thirty percent of search engine queries are based on the person names [1]. Nevertheless, extracting information about people from web search engines is a difficult task when a person is referred by different nick names.

For instance, a popular cinema artiste original name *Shivaji Rao Gaekwad* is referred by different alias names like “Super Star”, “Badsha”, ”Muthu”, “Robot”, “Chitti”, “Dancing Maharaja”, and much more. We will not be able to retrieve all information about the artiste from the web, unless we extract the top ranked alias names. Here, different entities can share the same name called lexical ambiguity.

On the other hand, a single entity can be designated by multiple names (i.e, referential ambiguity). A real-world example is alias name “Badsha” refers to *Shah Rukh Khan* another actor in the same domain of expertise. This problem is solved by semantic Meta data for entities and automatic extraction of Meta data [2] can accelerate the process of semantic annotation. For named entities, automatically extracted aliases can serve as a useful source of Meta data, thereby providing a means to disambiguate an entity. Identifying aliases of a name are important for extracting relations among entities. For example, Matsuo et al[3] propose a social network extraction algorithm in which they compute the strength of relation between two individuals X and Y by the web hits for the conjunctive query , “X “ and “Y”. However, both persons X and Y might also appear in their alias names in web contents. Consequently, by expanding the conjunctive query using aliases for the names, a social network extraction algorithm can accurately compute the strength of a relationship between two persons.

2. RESEARCH DIRECTIONS & EXISTING METHODS

Our research is headed towards building a web extraction system which extracts efficient patterns for Indian name aliases and further this system can be adapted to various fields.

Alias extraction is basically an information retrieval task [IR], which looks for similar, preceding, succeeding, adjacent, lexico-syntactic, supervised co-occurring text from a large cluster of documents. The main function of information retrieval is to build a term-weighting system [4] which will enhance the retrieval effectiveness. Two measures are normally used to assess the ability of a system to retrieve the relevant and reject the non-relevant items of a collection, which is known as Recall and Precision respectively. Determining recall and precision is the significant accuracy measure of any information retrieval task in web and holds good for alias extraction too.

Below we will discuss further on various techniques viz., Word Association Norms and lexicography, collocation extraction in natural language processing, cross-document co-reference resolution, Duplicate Detection Using learnable String Similarity Measures, Unsupervised clustering to identify the referents of personal names, self-annotating web, people searching strategies in World Wide Web, Disambiguating web appearances of people in a social network, ‘PolyPhonet’-a Social network system, disambiguating name sakes, approximate string matching method, mnemonic extraction, approximate name matching using finite state graphs, measuring semantic similarity between words, Weps-2 Evaluation campaign and Web mining for Alias extraction which has been used in this area.

2.1 Word Association Norms and Lexicography

In linguistics, it is a general practice to classify words not only on the basis of their meaning but also on the basis of their co-occurrence with other words. The word ’bank’ has dual meaning with respect to the association of adjacent words and expressions. For instance words such as, currency, cheque, loan, account, interest etc., are related with financial institutions. On the other hand, bank co-occurring with water, boat etc., are related to river. Word association norms are well known to be an important factor in psycholinguistic research, specifically in the area of lexical retrieval. People understand quicker than normal to

the word 'nurse' if it follows a frequently associated word such as 'doctor'. It is found in psycholinguistic research that the word 'doctor' is most often associated with 'nurse' followed by sick, health, medicine, hospital . In this paper, association ratio [18] was proposed for measuring word association norms from computer readable corpora based on information theoretic concept of mutual information.

2.1.1 Mutual Information:

Mutual Information states that if two points (words), x and y have probabilities p(x) and p(y), then their mutual information, I(x,y) is defined to be

$$I(x,y) = (\text{Log } P(x,y))/(P(x)\ p(y))$$

Informally, mutual information compares the probability of observing x and y together with the probabilities of observing x and y independently (chance). If there is a genuine association between x and y then the joint probability P(x,y) will be much larger than chance P(x) P(y), and consequently I(x,y) > 0. If there is no interesting relationship between x and y, then P(x,y) = P(x) P(y), and thus I(x,y)=0. If x and y are in complementary distribution, then P(x,y) will be much less than P(X) P(Y), forcing I(x,y) < 0. Word probabilities P(x) and P(y) are estimated by counting the number of observations of x and y in a corpus, f(x) and f(y) and normalizing by N, the size of corpus. For experimentation, corpora of different sizes were used. Joint probabilities, P(x,y) are estimated by counting the number of times that x is followed by y in a window of words, $f_w(x,y)$, and normalizing by N. The window size parameter used to look at different scales. Smaller window identified fixed expressions (idioms such as bread and butter), larger window sized highlight semantic concepts and other relationships.

Mean and Variance of the Separation between Word X and Word Y

Relation	Words X	Word Y	Separation	
			Mean	Variance
Fixed	Bread	butter	2.00	0.00
	Drink	drive	2.00	0.00
Compound	Computer	scientist	1.12	0.10
	United	States	0.98	0.14
Semantic	Man	woman	1.46	8.07
	Men	women	-0.12	13.08
Lexical	refraining	from	1.11	0.20
	Coming	from	0.83	2.89
	Keeping	from	2.14	5.53

From the above table, it was inferred that fixed expressions such as 'bread and butter' or 'drink and drive', the words are separated by a few numbers of words. They often found very close to each other within five words. Hence, mean separation is two, and variance is zero. Compound expressions also appear close to each other. In contrast, semantic words like man/woman have larger variance in their separation. Lexical relations come in several varieties. There are some like 'refraining from' are fairly fixed, others 'coming from' separated by an argument, and still others like 'keeping from' are almost certain to be separated by an argument. Technically association ratio is different from mutual information in two aspects. First, joint probabilities are supposed to be symmetric: P(x,y) =P(y,x) and thus mutual information is also symmetric: I(x,y) = I(y,x). However, an association ratio is not symmetric since f(x,y) encodes linear precedence. f(x,y) denotes the number of times that word x appears before y in the window of w words, not the number of times that the two word appears in either order. This work provides a precise statistical calculation that could be applied to a large corpus of text to produce a table of associations for tens of thousands of words. This association ratio could be an important tool to aid the lexicographer. It can help us decide what to look for; it provides a quick summary of what associated word must be in a readable corpora.

2.2 Extracting Collocations

Corpus analysis has been widely used by researchers after the tremendous growth rate of web. Corpus analysis extracts collocations by using automatic techniques for retrieving lexical information from textual corpora. Collocations refer to sequence of words that co-occur in a web document. Natural languages are full of collocations, recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages. Research work in lexicography indicates that collocations are common in all types of writing, including both technical and non-technical modes of communication. Xtract [5] software consists of a set of tools to locate words in context and make statistical observations to identify collocation in web documents. Xtract uses straight statistical measures to retrieve from a corpus pair-wise lexical relations whose common appearance within a single sentence are correlated. The advantage of Xtract is that it can be used to produce collocations involving more than two words (n-grams). Evaluation of any retrieval systems is usually done with two parameters precision and recall [Salton 1989] and it is evident that it is used to assess the quality of retrieved material. However, Xtract tool cannot be directly applied to extract aliases, since nick names of celebrities need not be a natural language collocation.

2.3 Cross-Document Co-Reference Problem

Cross-document co-reference [CDC] is another problem in alias extraction using natural language. Bagga and Baldwin et al [6] discovered a cross-document co-reference resolution algorithm which uses the vector space model (SVM) to resolve ambiguities between people having the same name. Initially, they developed a co-reference algorithm which works in two steps (1) Extract co-reference chains for within the document (2) Clustering co-reference chains under a SVM to identify all names mentioned in the document set. However, due to enormous documents on the web it is impractical to perform within-document co-reference resolution to each document separately and cluster the documents to find aliases. Moreover, the noise and different writing styles followed in web documents make it difficult to perform within-document co-reference resolution. Later, they devised a cross-document co-reference resolution which works across each document separately with a better accuracy.

2.4 Duplicate Detection Using Learnable String Similarity Measures

The problem of identifying approximate duplicate records in databases is an essential step for data cleaning and data integration processes. Duplicates can cause data-mining algorithms from discovering important regularities. This problem is typically handled during a tedious manual data cleaning, or "de-duping", process.

Previously they have been using manually tuned distance metrics for estimating the potential duplicates. The author presented two learnable text similarity measures[16] suitable for this task: First one uses the Expectation Maximisation (EM) algorithm for estimating the parameters of a generative model based on learnable string edit distance, and a novel vector-space based measure that employs a Support Vector Machine (SVM) to obtain a similarity estimate based on the vector-space model of text. The character based distance is best suited for shorter strings with minor variations, while the vector-spaced representation is more appropriate for fields contain longer strings with more variations. The overall duplicate detection system , MARLIN (Multiply Adaptive Record Linkage with INduction), employs a two-level learning approach, first string similarity measures are trained for every database field so that they can provide accurate estimates of string distance between values for that field. Next, a final predicate for detecting duplicate records is learned from similarity metrics applied to each of the individual fields. Utilized Support Vector machines for evaluation and showed that it outperforms previous methods such as decision trees, and classifiers [56,57]. It has been proved MARLIN can lead to improved duplicate detection accuracy over traditional techniques.

The Fig 1 is the framework for improving duplicate detection, using trainable measures of textual similarity and will provide significant value addition in alias extraction.

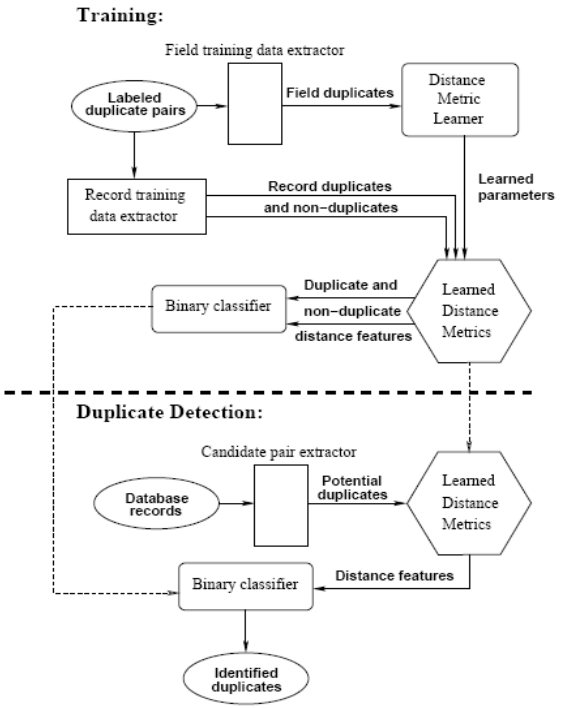

Fig 1 Duplicate Detection Framework [MARLIN]

2.5 Unsupervised Clustering to Identify the Referents of Personal Names

In this paper, a set of algorithms were described for disambiguating personal names [7] with multiple real referents in text, based on little or no supervision. The approach utilizes unsupervised clustering technique over a rich feature space of biographic facts, which are automatically extracted via a language-independent bootstrapping process. The induced clustering of named entities are then partitioned and linked to their referents via the extracted biographic data.

One open problem in natural language ambiguity resolution is the task of proper noun disambiguation. While word senses and translation ambiguities may typically have 2-20 alternative meanings that must be resolved through context, a personal name such as "Bill Clinton" may potentially refer to hundreds or thousands of distinct individuals.

Supposing, the Search of Google shows 30 web pages mentioning "Bill Clinton", of which the top 5 unique referents are: Bill Clinton – Professor at University of Malaysia, Former President of USA, Film producer in Hyderabad, Gun Dealer in New Delhi, a Computer science student in Japan.

Each different referent typically has some distinct contextual characteristics. These characteristics can help distinguish, resolve and trace the referents when the names appear in online documents.

2.6 The Self-Annotating Web

The growth of Semantic Web depends on the availability of ontology's as well as on the large number of web pages annotated with Meta data conforming to these ontologies. Previous ontology structures being defined in research projects like DAML, OASIS, and UDDI have problems related to missing Meta data. The major issues in those ontology structures were laborious task requiring a lot of time and expert know-how.

The principal idea of the self-annotating web [2] (PANKOW-Pattern Based Annotation through Knowledge On the web) is that it uses globally available Web data and structures to annotate local resources semantically to bootstrap the Semantic Web. A method employs an unsupervised, pattern-based approach to categorize instances with regard to a given ontology. The approach is novel combining the idea of using the linguistic patterns to identify certain ontological relations as well as the idea of using the web as a big corpus to overcome lack of meta data. It is unsupervised as it does not rely on any training data annotated by hand and it is pattern based in the sense that it makes use of linguistically motivated regular expressions to identify instance-concept relations in text. The driving principle of PANKOW is one of disambiguation by maximal evidence in the sense that for a given instance it proposes the concept with the maximal evidence derived from Web statistics. PANKOW has been conceived for our annotation framework CREAM [38] and implemented on Onto Mat using queries to the web service API of Google. The automatic annotation produced by PANKOW has been evaluated against semantic annotations produced by two independent human subjects.

2 .7 People Searching strategies in World Wide Web

This paper describes the creation of a Test Bed [12] to evaluate people searching strategies on the World Wide Web. The task involves resolving person names ambiguity and locating relevant information characterizing every individual under the same name. Person names are highly ambiguous: For Instance, only 90,000 different names are shared by 100 million people according to the U.S Census Bureau [1]. Given an ambiguous name as query in a search engine, disambiguation algorithm groups all documents related to the given ambiguous name in to a single cluster. In this paper, a test bed created to evaluate strategies addressing the people searching task on web documents. (i) a corpus of web pages retrieved using person names as queries to web search engines (ii) a classification of pages according to the different people they refer to (iii) manual annotations of relevant information found in the web pages describing them such as email, image, profession, phone number, etc, (iv) the results of applying a general purpose clustering algorithm to that annotated data, which serve as a baseline for the ambiguity resolution problem. Weps corpus test bed were created in the following steps (1) Generating ten English person names, using random combinations of the most frequent first and last names in the U.S Census 1990. (2) Collecting the first 100 pages retrieved by the Google Search engine for every specific person name. (3) Grouping documents according to the person they refer to , for every person name (4) Classifying every web document in the collection as a (i) home page entry (ii) part of a home page (iii) reference page (exclusively containing the information about a person) (5). Annotating all the occurrences of certain types of descriptive information such as name, job, person image, date of birth, date of death, email, postal address, fax number, phone number, location, author of (e.g books, paintings, patents, …) if any and other descriptions. It was implemented and tested using Agglomerative Vector Space Clustering algorithm, which has been previously used to evaluate the similar task [55]. Future directions are to expand and balance the corpus, including two additional types of person names: less frequent names for which ambiguity must certainly be lower. On the other hand, for a 'Celebrity Person', the ambiguity be certainly be higher and it is anticipated to dominate the top-ranked results of search engine results.

2.8 Disambiguating Web Appearances of People in a Social Network

A search engine returns many pages for that person's name but which pages are about the person you care about, and which are about other people who happen to have the same name? Furthermore, if we are looking for multiple people who are related in some way, how can we best leverage this social network? We present two unsupervised frameworks for solving this problem: one based on link structure of the Web pages, another using Agglomerative/Conglomerative Double Clustering (A/CDC)—an application of a recently introduced multi-way distributional clusters method. To evaluate our methods, we collected and hand-labeled a dataset of over 1000 Web pages retrieved from Google queries on 12 personal names appearing together in someone's in an email folder. On this dataset our methods outperform traditional agglomerative clustering by more than 20%, achieving over 80% F-measure.

2.9 Social Network System

Social network plays important role in the semantic web viz., Information Retrieval, Knowledge Management, and Ubiquitous computing and so on. People conduct communications and share information through social relations with other such as friends, family, Class mates, colleagues, collaborators, and Business partners. Social networking services have gained popularity in the recent years. SNS's are useful to register personal information including a user's friends and acquaintances on these systems; the systems promote information exchange such as sending messages and reading weblogs, *Friendster*, *Orkut*, *Face book*, *Twitter* are the successful SNS. In the context of semantic Web, social networks are crucial to realize a web of trust, which enables the estimation of information credibility and trustworthiness [40]. Because anyone can say anything on the Web, the Web of Trust helps humans and machines to discern which contents are credible and to determine which information can be used reliably. Ontology construction is also related to social network. *Kautz and Selman et al* developed a social network extraction system from the Web called Referral Web [41]. The system focuses on co-occurrences of names on Web pages using a search engine. It estimates the strength of relevance of two persons X and Y putting a query "X and Y" to a search engine. If X and Y have strong relation, we can find much evidence with their homepages, list of co-authors in technical papers, and organizational charts. A path from person to person is obtained automatically. Later, with the development of WWW and semantic Web technology, more information on our daily activities has become online. Due to greater potential and demand this method seems to be outdated. *P.Mika et al* developed a system for extraction, aggregation and visualization of online social networks for a semantic web community, called Flink [42]. Social networks are obtained using analyses of Web pages, e-mail messages, and publications and self created profiles (FOAF Files). The Web mining component Flink also employs a co-occurrence analysis. Given a set of names as input, the component uses a search engine to obtain hit counts for individual names as well as the co-occurrence of those two names. The system targets semantic web community. Therefore the term, "semantic web OR ontology" is added to the query for disambiguation.

McCallum et al and his group [43] present an end-to-end system that extracts a user's social network. The system identifies unique people in e-mail messages, finds their homepages, and fills the fields of a contact address book as well as the other person's name. Links are placed in the social network between the owner of the web page and persons discovered on that page. *Harada et al* [44] develop a system to extract names and also person-to-person relations from the Web. *Faloutsos et* al obtain a social network of 15 million persons from 500 million Web pages using their co-occurrence within a window of 10 words. *Knees et al*[45] classify artists into genres using co-occurrence of names and keywords of music in the top 50 pages retrieved by search engine. *L.Adamic et al* classified

the social network of Stanford university students, and collected relations among students from Web link structure and text information.

In this paper, an advanced social network extraction system called PolyPhonet [3] were introduced, which employs advanced techniques to extract relations of persons, detect groups of persons, and obtain keywords for a person. It is a Web-based system for an academic community to facilitate communication and mutual understanding based on a social network extracted from the Web. The system has been used at JSAI annual conferences for three years and at UbiComp2005. Person names co-occur with many words on the Web. A particular researcher's name will co-occur with many words that are related to that person's major research topic. This paper uses person-to person matrix called adjacent matrix and person-word co-occurrence matrix as affiliation matrix. The multi-faceted retrieval is possible on the social network: researchers can be sought by name, affiliation, keyword, and research field, related researchers to retrieved researcher are listed; and a search for the shortest path between two researchers can be made. We can measure the similarity of two research paper contexts. In the Researcher's cases, we can measure how mutually relevant the two researcher's research topics are: if two persons are researchers of very similar topics, the distribution of word co-occurrences will also be similar.

Even more complicated task such as searching for a researcher who is nearest to a user on the social network among researchers in a certain field. 'PolyPhonet' is incorporated with a scheduling support system [46] and a location information display system [47] in the ubiquitous computing environment.

Google is used to measure co-occurrence of information and obtain web documents.

2.10 Mining the Web for Mnemonic Name Extraction

The web is a source from which we can collect and summarize information about a particular real-world object. The proliferation of tools like bulletin boards and web logs initiated the need for information extraction, disseminating knowledge and much work has been done in this area. Major problem is that the same object is referred in different ways in different documents. For example, a person may be referred to by full name, first name, affiliation and title or nick names. The term mnemonic name refers to an unofficial name of object. Generally, people use nick names or mnemonic names when they complain or evaluate an object unfavourably. Here, full name of a person is considered as official name.

The first novel method for extracting mnemonic names from the web is proposed by Hokama and Kitagawa et al[10] [Fig 2].Evaluative expressions about an object (e.g business organization, product, person) are extracted from text surroundings the string that represents that object. The ability to collect web pages describing the target object is first needed to extract representing information. Existing research extracts evaluative expression from text surrounding the official name of target object, such as product name. Specialized topic detection for a particular object will become important as well as reputation information extraction, which analyses text around the object's name and extracts "local information". Personal information sources exist, personal databases, public home pages, Wikipedia. These are static or official, so they cannot yield dynamic and unofficial information that includes recent popular topics about a person. For larger web space, information sources such as bulletin boards and blogs must be tapped to collect dynamic and unofficial sources. We need to know how much attention these topics attract to the public. In this method, short strings adjacent to the full name of target person to extract mnemonic names. This method is applicable only to extract Japanese texts, because it uses a Japanese linguistic language. Object identification or name entity recognition aims to discover official names of entities; purpose is to extract "non-official" names of people.

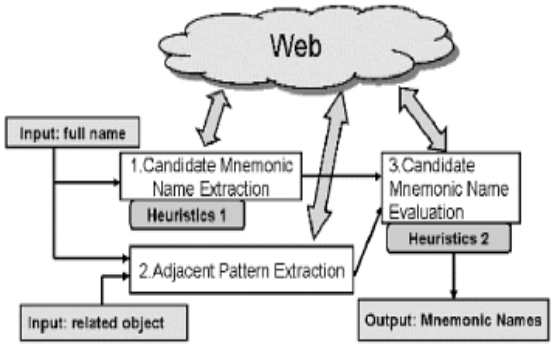

Fig 2 Extracting Mnemonic names of people from the web

This method comprises of three components: 1) Extracting candidate mnemonic of target person from web 2) Extracting string adjacent to the first name of target person using prefix and suffixes. 3) Evaluate candidate mnemonic names extracted in step1, using adjacent patterns extracted in step 2. Then, select top k candidates as mnemonic of target person. For a given name p, they search for the query "* koto p" and extract the content that matches the asterisk, Koto "<<string in Japanese>>" in English which is equivalent to "be called", and also it is a vague term. It can be a clue for searching but not the decisive factor. The Japanese language word koto has multiple meanings "also known as", "incident", "thing", "matter", "experience", and "task".

Candidate Mnemonic Name Extraction

The pattern "alias<<Japanese string>>full name" in Japanese language is used to describe the alias name or nick names of person. Therefore, string that occurs right before the "<<Japanese string>>full name" is a mnemonic name. If this string is commonly used as mnemonic name, it occurs in web repeatedly, based on this, we extract candidate mnemonic names.

1. Perform a query "<<Japanese>>full name" on a web search engine then get the URL list
2. Get web pages in URL list and analyse these pages, thus extract the string <t1 t2 t3....tn> that occurs right before the string "<<Japanese string>>full name" (t1t2.....tn) are morphemes
3. Extract sub-strings of <t1 t2 ....tn> and then select sub-strings when first morphemes POS tag is a "general noun" as candidate mnemonic names.

Then count frequency of occurrence for each candidate.

4. Eliminate candidate mnemonic names that occur only once in analysed web pages.

Adjacent Pattern Extraction

We get web pages including the full name of target person by performing a web search, and extract strings adjacent to the full name i.e. prefix and suffix patterns. Because, there can be people with same name, we add an object name that has relevance to the person (i.e. parent organization) to the search query. After extracting all prefix and suffix patterns, we calculate weights for all patterns by considering the co-occurrence relation between full name and patterns.

The procedure for adjacent pattern extraction follows

1. Let the object name that has great relevance to the target person be rel-object. Perform the query "Full Name AND rel-object" on a web search engine, then get URL list
2. Analyse web pages in the URL list and extract strings adjacent to the full name<t1 t2 t3.....tm> extract sub-strings of <t1 t2, t3.....tm> in a similar way of candidate mnemonic names extraction and add these sub-strings to the list of prefix and suffix patterns.
3. Calculate weights for all extracted prefix patterns as follows. The weight of a prefix pattern, w(prefix) is calculated by the formula

   R=searchResults(Prefix)
   R1=searchResults("PrefixFullName")
   W(prefix)=R1/R
   R1 - refers to number of web pages including "prefix- Full Name"
   R - refers to number of web pages including Prefix

   searchResults(query) is a function that returns the total number of web pages including query. It is not possible to know the total number of those pages. This paper uses yahoo! API's , to get totalResultsAvailable Field for the estimated value.
4. Calculate weights for all extracted suffix patterns as follows. The weight suffix pattern w(suffix) is calculated by the same formula

   R=searchResults(suffix)
   R1=search Results("FullNamesuffix")
   W(suffix)=R1/R
5. Add prefix and suffix pattern whose weights exceed the given threshold to the "adjacent patterns list"

Candidate Mnemonic Name Evaluation

Final step evaluate extracted mnemonic names using second heuristics. It is possible that a candidate mnemonic name *cand* is actually a mnemonic name of target person if *cand* occurs just before or just after adjacent pattern.

The evaluation procedure as follows:

1. Set the initial score of *cand* as 0
2. For all adjacent patterns, apply this procedure

   (a)If the adjacent pattern is prefix pattern generate string "*prefix cand* "

   If the adjacent pattern is suffix pattern, generate a string "*cand suffix*"

   (b) Obtain the total number of web pages including the generated string total used by a search engine. Add the product of total and the pattern's weight (w (prefix)) or w (suffix)) to the score.

This calculation because it is highly possible that *cand* is the actual mnemonic name in those situations; there are many web pages including "prefix *cand*" or "*cand suffix*" and the pattern's weight is big. After calculating all scores of candidate mnemonic names, we select top k candidates as mnemonic names of the target person. Hence, this method works well for extracting mnemonic names in a Japanese websites. Results returned by the method for six different people were proved to be correct.

However, there are few inappropriate mnemonic names and a few appropriate mnemonic names were missing. In spite of, two post processing heuristics specific for Japanese language to filter out incorrect mnemonic names, the method seem to produce incorrect retrieval under some cases. Moreover, Due to multiple meanings of said pattern many noisy and incorrect aliases were extracted. Future directions given by the author are improving pattern's weight and candidate's score calculation, investigating generality and robustness of the method, and extending the work to other objects and languages.

2.11 Approximate Name Matching Using Finite State Graphs

Some of the string matching algorithms [9] used for extracting variants or abbreviations of personal names. For instance, matching "Ram Kumar" with the first name initialized variant 'R. Kumar'. Approximate string matching is an interrelated area of natural language processing, Information Extraction, and Information Retrieval. Personal name can be considered as object tags that may appear in many different forms called as variants. A personal variant can be described as a text occurrence that is conceptually well related with the correct form or canonical form of a name. The recognition [Thomson and Dozier 1999] of variant of these sequences belongs to three categories: name-recognition, name-matching and name searching.

Another method for common name extraction and there problems, Name Recognition is the process by which a string of characters is identified as a name. It is widely used to extract names from texts as described in Message Understanding Conferences [MUC-4,1992;MUC-6,1995] as a part of information extraction. In [MUC6,1995], the recognition of the named entity is considered as a key element of extraction systems. Entities include names of persons, organizations, or places as well as expressions of time or monetary expressions. In [MUC-7,1997], the named entity recognition implies identifying and categorizing three subareas which are the recognition of time expressions, numerical expressions, and entity names-person, organizations, and places. Name matching

corresponds to determining whether two strings of characters previously recognized as names actually designate the same person. Name matching does not focus on the case of various individuals who have identical name labels. In this case, two possibilities arise (1) The matching is exact, there is no problem (2) matching is not exact, making it necessary to determine the origin of these variants and apply approximate string matching.

Name–Searching designates the process through which a name is used as part of a query to retrieve information associated with that sequence in a database. Here, two problems can appear. (1) The names are not identified as such in the data base registers or in the syntax of the query, Name recognition techniques are needed. (2) The names are recognized in the database records and in the query, but it is not certain that the recognized names designate the same person. It does not require any matching techniques. In bibliographic databases and citation index systems, variant forms create problems of inaccuracy which affects information retrieval. It means ultimately it affects the quality of information from databases and the citation statistics used for the evaluation of scientist work. A number of string matching techniques had been developed to validate variant forms, based on similarity and equivalence relations. This variant identification requires binary matrices and finite-state graphs. This procedure was tested on samples of author names from bibliographic records, Library and Information Science Abstracts and Science Citation Index and Expanded databases. The evaluation includes precision and recall as a proof for completeness and accuracy. However, an inherent limitation of such string matching methods would not identify aliases.

2.12 Measuring Semantic Similarity

Semantic similarity measures play significant roles in information retrieval and Natural Language Processing. Semantic similarity measures [19] were used in automatic query suggestion and expansion. Previous work used the same for community mining, relation extraction, automatic meta data extraction. Semantic similarity between entities changes over time and across domains. For example, a user may be interested to retrieve information about "apple" in the sense of apple computer and not apple as a fruit. New words are constantly being created as well as new senses are assigned to existing words. This paper proposes an automatic method to measure semantic similarity between words or entities using web search engines. Page counts and snippets are two useful information sources provided by most search engines. Page count for the query "P AND Q" can be considered as a global measure of co-occurrence of words P and Q. For example, page count of the query "apple" AND "computer" is higher than "banana" AND "computer". The page counts for the former query is higher and it indicates that "apple" is more semantically similar to computer than "banana". However, there are some drawbacks. Page count need not be equal to the word frequency because the queried word might appear many times on one page. Moreover, page count of a polysemous word [multiple senses] might contain a combination of all its senses. For these reasons, page counts alone are unreliable measure for semantic similarity. Here, a new method that considers both page counts and lexico-syntactic patterns extracted from snippets to overcome the above problem.

2.12.1 Lexico-syntactic patterns:

Consider the following snippet from Google for the query 'Jaguar and Cat'

"The Jaguar is the largest Cat in western hemisphere usually found in forests"

Here, the phrase *is the largest* indicates a hyponym relationship between the Jaguar and the Cat. Phrases such as "also known as", "is a part of", "is an example of" all indicate various semantic relations of different types. Such indicative phrases have been applied to various tasks with better results, such as hyponym extraction [25] and fact extraction [Pasca et al. 2006]. From this example, we can say pattern X *is the largest* Y, where we replace the two words Jaguar and Cat by two wild cards X and Y. Therefore, an automatic extraction of lexico-syntactic pattern based approach was proposed and semantic similarity had been proved using text snippets obtained from a web search engine. Also, it is integrated with web-based similarity measures using WordNet, Synsets and support vector machines to create a robust semantic similarity measure. This integrated method has been proved with a existing benchmark dataset. And also this was the first attempt to combine WordNet, Synsets and web content to leverage a robust semantic similarity measure. Proven method was tested in a community mining task in capturing similarity between real-world entities, and also proved to be worthy in word sense disambiguation [WSD Mc Carthy et al 2004].

2.12.2 Word Sense Disambiguation

Contextual Hypothesis for Sense [Schutze 1998] states that the context in which a word appears can be used to determine its sense. For instance, a web page discussing 'Jaguar as a car', is likely to narrate about other types of cars, parts of cars etc., whereas a web page on 'Jaguar the cat' is likely to contain information about other types of cats and animals. Paper computed precision, recall and F-score for each cluster created for the top 1000 snippets returned by Google for two ambiguous entities Jaguar and Java. Jaguar can have any one of three senses with respect to the context like cat, car or an operating system. Ambiguous word Java also can have three senses like programming language, Island and coffee.

The scope of this work can be well utilised in automatic synonym extraction, query suggestion and finally name alias recognition.

2.13 Web People Search Clustering Task Evaluation Campaign-Weps2

The second Web people Search Task (WEPS2)[13] evaluation campaign took place with participation of 19 research groups around the world. Given the output of a Web search engine for an ambiguous person name as a query, two tasks were to be resolved. 1) A clustering task to group related web pages of the same person 2) Extraction task which consists of extracting salient attributes for each persons sharing the same name. This paper presents resources, methodology and evaluation metrics, participation and comparative results for the clustering task.

When a single name is shared by many people it becomes a challenging task. This ambiguity has recently become an active research topic and simultaneously a relevant application domain for Web search services. Websites like zoominfo.com, Spock.com, 123people.com perform web people search, with fair disambiguation capabilities. As the amount of information in WWW grows, more of these people are mentioned in different web pages. Therefore, query for a common name in the Web will usually produce a list of results where different people are mentioned. In order to retrieve the right name what the user is interested in, user might refine the original query with additional terms to filter out irrelevant one. In some cases, the existence of a pre-dominant person such as a celebrity makes it likely to dominate the ranking of search results, complicating the task of finding information about other people sharing the same name. The disambiguation of person names in Web results are compared to other tasks such as Word Sense Disambiguation (WSD) [1] and Cross-Document Co-reference (CDC) [6]. It is useful to point the crucial differences between WSD, CDC and Web People Search

1) WSD typically concentrates in the disambiguation of common words (nouns, verbs, adjectives) for which is relatively small number os senses exist. Word senses in dictionaries often have subtle differnces.

2) WSD can rely on dictionaries to define the number of possible senses for a word. In the case of name ambiguity no such dictionary is available.

3) The objective of CDC is to reconstruct the co-reference chain for every mention of a person. In Web person name dismbiguation it suffices to group the documents that contain at least one mention to the same person.

The first Weps-1 Evaluation [49] gathered 17 research teams. The task was to cluster search results for a given name according to the different people that share his name. A large dataset was collected and manually annotated. This research becomes the de-facto standard for the disambiguation task and used beyond Weps [50,51,52,53].

2.14 Mining the Web for Alias Extraction

In the second alias extraction method Danushka Bollegala et al [11] [35], discovered a novel approach to find aliases of a given name from the web. Method (Fig 3) comprises of two components: lexical pattern extraction, and candidate alias extraction and ranking. Exploited a set of known name and their aliases (name-alias pair) as training data to generate lexical patterns that convey information related to aliases of names from text snippets returned by a web search engine. Since web contents are dynamic in nature, using the initial seed search engine retrieves text snippets as available in the web documents.

The training data is allowed to learn for every new search of a lexical-pattern and it is updated every time. Different combination of pattern is given to search engine for maximizing retrieval of lexical-patterns relevant to the given personal name. The patterns are then used to find candidate aliases of a given name. Anchor texts and hyperlinks were used to design a word co-occurrence model and define numerous ranking scores to evaluate association between a name and its candidate aliases.

2.14.1 Lexical Pattern Extraction

Search engine plays as a source in name alias extraction. It provides a brief snippet for each search result by selecting portion of text that appears in web page within the proximity of query. Such snippets provide valuable information related to the local context of query. For example, consider the snippet returned by Google for the query “Will Smith * The Fresh Prince”.

Here, wild card operator * is used to perform a NEAR query and it matches with one or more words in a snippet. The snippets are parsed with various patterns. This Lexica-syntactic pattern approach have been used in numerous related tasks such as extracting synonyms, hyponyms [25] and metonyms [26].

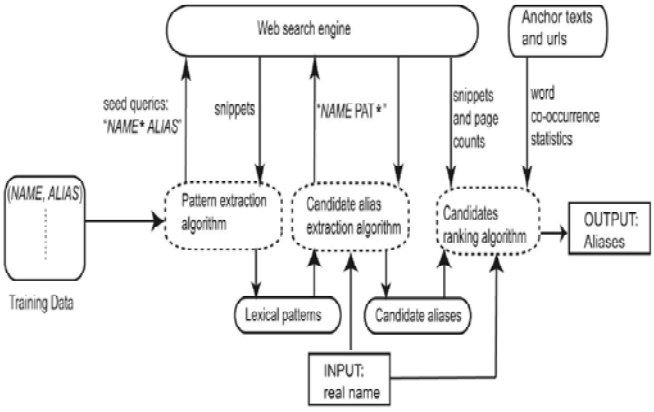

Fig 3 Personal Name Alias Extraction of Celebrities from web

The above diagram Fig 3, explains the extracting approach of alias. First a set with Name + alias will be extracted from the web Fig 3. Then, from each snippet, the Create Pattern function extracts the sequence of words that appear between the name and the alias. We repeat the process described above for the reversed query, “ALIAS *NAME” to extract patterns in which alias precedes the name. Like-wise there are another eight patterns [Fig 4] given in the query to extract snippets from web documents, each input pattern is evaluated for performance with a measure called F-Score.

?? Muthu(1995)[Actor **Rajnikanth**] ... **aka** "The **Dancing Maharaja**" - India *(English title)*

Fig 4. A snippet returned for the query “Rajnikanth, aka the “Dancing Maharajah” by Google

2.14.2 Candidates Ranking

Considering the noise in web snippets, candidates extracted by shallow lexical patterns might include some invalid aliases. Among these candidates we should identify candidates which are most likely to be correct aliases of a given name. Alias ranking is necessary to identify which one of the candidate alias is most likely to be correct alias in the alias set. Here, candidate ranking is done with ranking scores to measure the association between a name

and a candidate alias using three different approaches: (1) lexical pattern frequency (2) Word Co-occurrences in an anchor text graph, and (3) Page counts on the web.

ExtractPattern algorithm extracts over 8000 patterns for 50 English personal names in the data set used. However, not all patterns are equally informative about aliases of a real name. Input query patterns are ranked according to their F-Score to identify top k efficient patterns for study (fig.5). F-Score of a pattern's' is computed as the harmonic mean between the precision and recall of the pattern. The following table shows, ranking of different patterns given in query for generating lexical pattern output using English personal name dataset.

Pattern	F-score
* aka [NAME]	0.335
[NAME] aka *	0.322
[NAME] better known as *	0.310
[NAME] alias *	0.286
[NAME] also known as *	0.281
* nee [NAME]	0.225
[NAME] nickname *	0.224
* whose real name is [NAME]	0.205
[NAME] aka the *	0.187
* was born [NAME]	0.153

Fig 4: Top Lexical patterns with F-score measure

F-Score improves as a result of the improvement in recall. This algorithm is ideally suited to extract patterns written in languages other than English. Results had been proved that it was quite successful in Italian and French languages, and patterns that contain punctuation symbols also appear among the top 200 patterns used for measuring the completeness and accuracy of this work.

To measure the strength of association between a name and a candidate alias the following nine popular statistical measures were used to perform ranking of candidate aliases. Statistical measures are Co-occurrence frequency(CF), Term-frequency-inverse document frequency(tfidf), chi-squared measure(CS), Log-likelihood ratio(LLR), Point-wise mutual information(PMI), Hyper-geometric distribution(HG), Cosine measure (cosine), Overlap, and Dice co-efficient. The nine statistical ranking scores are integrated in to a single ranking function using ranking support vector machines (SVM). Paper used linear, quadratic, and radial basis function (RBF) kernels for ranking SVM. The Statistically significant Mean Reciprocal Rank (MRR) and Average Procision (AP[22]) is used to evaluate the different approaches. The MRR of this method is 0.67 yielded better performance than previous method of Hokama and Kitagawa et al[10] .

AREAS OF INTEREST

The name extraction approach using lexical pattern can be extended to medical, bio-medical, product and domain classification. Alias extraction can be used as a tool for both legitimate and illegitimate web applications. Also, virtual email addresses could be traced out.

We also focused our research in using regional markers like "patta peyar", "punai peyar" etc., for alias extraction but the results were not compromising. Since English is more dominant in web world.

CONCLUSION

This paper we have shown the existing algorithms there challenges and various methods which was proposed earlier.

The proposed work is foreseen to have significant advantages in web extraction. The extent this algorithm can be used will not be limited. In future, alias extraction can become a common tool for 'non-celebrities' and further would be extended to support different languages. Also, it can be used in inter-disciplinary fields for retrieving need based information from readable corpora.

REFERENCES

[1] R.Guha and Garg, "Disambiguating People in Search", Technical Report, Stanford University., 2004

[2] P.Cimano S.Handshuh, and S.Staab,"Towards the Self Annotating Web," Proc. Int'l World Wide Web Conf.(WWW'2004)

[3] Y.Matsuo, J.Mori, M.Hamasaki, K.Ishida, T.Nishimura,H.Takeda,K.Hasida, and M.Ishizuka, "Polyphonet:An advanced Social Network Extraction System,"Proc.WWW' 2006

[4] G.Salton and C.Buckley, "Term-Weighting Approaches in Automatic Text Retrieval,"Information Processing and Management" , 1988

[5] F.Smadja, "Retrieving Collocations from Text:Xtract" Computational Linguistics, 1993

[6] A.Bagga and B.Baldwin, "Entity-Based Cross-Document Co-referencing using the Vector Space Model," Proc. Int'l conf. Computational Linguistics, 1998

[7] G.Mann and D.Yarowsky, "Unsupervised Personal Name Disambiguation," Proc.Conf. Computational Natural Language Learning, 2003

[8] R.Bekkerman and A.McCallum, "Disambiguating Web Appearances of People in a Social Network," Proc. Int'l World Wide Web Conf, 2005

[9] C.Galvez and F.Moya Anegon, "Approximate Personal Name-Matching through Finite State Graphs,"Journal of American Society for Information Science and Technology, 2007

[10] T.Hokama and Kitagawa , "Extracting Mnemonic names of People from the Web," Proc. Ninth Int'l conf.Asian Digital Libraries, 2006

[11] Danushka Bollegala, Taiki Honma, Yutaka Matsuo and Mitsuru Ishizuka, "Mining for personal Name Aliases on the Web", In proc. of WWW '2008

[12] J.Artiles, J.Gonzalo, and F.Verdejo, "A Testbed for People Searching Strategies in the WWW," Proc.SIGIR'05, 2005

[13] S.Sekine and J.Artiles, "Weps 2 Evaluation Campaign:Overview of the Web People Search Attribute:Extraction Task," Proc. second Web People Search Evaluation Workshop (WePs '09) at 18th Int'l World Wide Web conf,. 2009

[14] G.Salton and M.McGill, "Introduction to Modern Information Retreival", Mc Graw-Hill 1986

[15] M.Mitra, A.Singhal, and C.Buckley, "Improving Automatic Query Expansion," Proc. SIGIR '98 1998

[16] M.Bilenko and R.Mooney, "Adaptive Duplicate Detection using Learnable String Similarity Measures," Proc. SIGKDD'03, 2003

[17] C.Manning and H.Schutz, "Foundations of Statistical Natural Language Processing", MIT Press, 1999

[18] K.Church and P.Hanks, "Word Association Norms, Mutual Information and Lexicography," Computational Linguistics 1991

[19] D.Bollegala, Y.Matsuo and M.Ishizuka, "Measuring Semantic Similarity between Words using Web Search Engines," Proc. Int'l World Wide Web Conf', 2007

[20] T.Joachims, "Optimizing Search Engines using Clickthrough Data," Proc. ACM SIGKDD'02, 2002

[21] T.Kudo, K.Yamamoto, and Y.Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis," Proc. Conf.Empirical Methods in Natural Language (EMNLP '04), 2004

[22] R.Beaza-Yates and B.Riberio-Neto, "Modern Information Retrieval", ACM Press,1999

[23] P.Mika, "Ontologies Are Us:A Unified Model of Social Networks and Semantics," Proc. Int'l Semantic Web Conf.(ISWC'05), 2005

[24] T.Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," Computational Linguistics, 1993

[25] M.Hearst, "Automatic Acquisition of Hyponyms from large Text Corpora," Proc.Int'l conf. Computational Linguistics,1992

[26] M.Berland and E.Charniak, "Finding Parts in Very Large Corpora," Proc.Annual Meeting of the Association for Computational Linguistics, 1999

[27] S.Chakrabarti, "Mining the Web: Discovering knowledge form Hypertext Data", Morgan Kaufman, 2003

[28] Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka, "Identifying People on the Web through Automatically Extracted Key phrases", In proc.of WWW

[29] G.Salton A.Wong and C.S.Yang, "A vector space model for Information Retrieval", Communications of the ACM, 1975

[30] G.D.M.Renmie and T.Jaakkola, "Using term informativeness for named entity detection," In por.c of ACM SIGIR'05, 2005

[31] D.Bollegala, Y.Matsuo, and M.Ishizuka, "Extracting key phrases to disambiguate personal names on the web", In proc. CICLing 2006.

[32] Y.Li.Zuhair, A.Bandar, D.M, "An approach for measuring semantic similarity between words using multiple information sources", IEEE Transactions on Knowledge and Data Engineering, 2003

[33] D.Beeferman and A.Berger," Agglomerative clustering of a search engine Query log", In ACM SIGKDD, International conference on Knowledge Discovery and Data Mining (KDD), 2000

[34] Dekang Lin, "Automatic retrieval and clustering of similar words", 17th Int'l conf. On computational Linguistics, 1998

[35] Danushka Bollegala , Yutaka Matsuo , Mitsuru Ishizuka, "Automatic Discovery of Personal Name Aliases from the Web", IEEE Transactions On Knowledge and Data Engineering, 2011

[36] Gonenc Ercan, Ilyas Cicekli, "Using Lexical Chains for Keyword Extraction".

[37] Tarique Anwar, Muhammed Abulaish and Khaled Alghathbar, "Web Mining for Alias Identification: a First Step towards Suspect Tracking"

[38] S.Handshuh and S.Staab, "Authoring and annotation of web pages in CREAM. Int'l Proceedings of the 11th International World Wide Web Conference, WWW 2002, Honolulu, Hawaii, ACM 2002

[39] E.Agirre, O.Ansa, E.Hovy, and D.Martinez, "Enriching Very Large Ontologies using the WWW. In. Proc. Of the first workshop on Ontology Learning 2000

[40] J.Golbeck and J.Hendler, "Accuracy of metrics for inferring trust and reputation in semantic web-based social networks", In Proc. EKAW 2004

[41] H.Kautz , B.Selman, and M.Shah, "The Hidden Web", AI Magazine, 1997

[42] P.Mika, "Flink: Semantic Web technology for the extraction and analysis of social networks", Journal of Web Semantics, 2005

[43] A.Culotta, R.Bekkerman, and A.McCallum," Extracting social networks and contact information from email and the web". In. CEAS- 1, 2004

[44] M.Harada, Sh.Sato, and K.Kazama, "Finding authoritative people from the Web", In.Proc. Joint Conference Digital Libraries(JCDL2004), 2004

[45] P.Knees, E.Pampalk, and G.Widmer, "Artist Classification with Web-based data", In. 5th International Conference on Music Information Retrieval(ISMIR), 2004

[46] M.Hamasaki, H.Takeda, I.Ohmukai, and R.Ichise, "Scheduling support system for academic conferences based on interpersonal networks. In.Proc.ACM Hypertext 2004.

[47] T.Nishimura, Y.Nakamura, H.Itoh, and H.Nakamura, "System design of event space information support utilizing" In.Proc.IEEE ICDCS2004

[48] P.Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proc. Assoc. for Cmputational Linguistics (ACL) 2002

[49] J.Artiles, J.Gonzalo, and S.Sekine. The smeval-2007 Weps evaluation:Establishing a benchmark for the Web people search task. In. Proceedings of the Fourth International Workshop on Semantic Evaluations, ACL 2007

[50] D.Kalashnikov, R.Nuray-Turan, and S.Mehrotra, "Towards breaking the quality curse. A Web querying approach to web people search", In Proc. Of Annual International ACM SIGIR Conference, Singapore, July 2008.

[51] Z.Kozareva, R.Moraliyski, and G.Dias, "Web people search with Domain Ranking", In TSD '08: Proccedings of the 11th Internationnal conference on Text, Speech and Dialogue, 2008.

[52] H.Saggion, "Experiments on semantic-based clustering for cross-document coreference", In International Joint Conference on Naural Language Processing, 2008

[53] M.Sanderson , "Ambiguous Queries:Test collections need more sense",In SIGIR '08"Proccedings of 31st conference on Research and Development in Information Retrieval, USA 2008, ACM

[54] T.Hisamitsu amd Y.Niwa, "Topic-Word Selection Based on Combinatorial Probability", Proc. Natural Language Processing (NLPRS '01) 2001

[55] C.H.Gooi and J.Allan, "Cross-Document Co-reference on a Large scale Corpus", Technical report, Centre for Intelligent Information Retrival, Department of Computer Science, University of Massachusetts, 2004

[56] S.Sarawagi and A.Bhamidipaty, "Interactive de-duplication using active learning", In Proc. Of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2002), Edmonton, Alberta 2002.

[57] S.Tejada, C.A.Knoblock, and S.Minton, "Learning doimain-independent string transformation weights for high accuracy object identification. In.Proc. of the Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining, 2002

[58] Huang Xuanjing, Wu Lide, "Language-independent Text Categorization," 2000 International Conference on Multilingual Information Processing, pp 37-43, 2000

[59] Yiming Yang, Xin Liu, "A re-examination of text categorization methods," Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 1999, pp 42-49

[60] Tao Jiang, Ke Wang/Ah-Hwee Tan "Mining Generalized Assocaitions of Semantic Relations from Textual Web Content", IEEE Transactions on Knowledge and Data Engineering, Volume 19 Issue 2, February 2007

[61] Jian Zhang, Yiming Yang, "Robustness of regularized linear classification methods in text categorization," In Proceedings of SIGIR 2003. The Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[62] Y.Yao, "Measuring retrieval effectiveness based on user preference of documents", Journal of the American Society for Information science, 1995

[63] T.Joachims, "Learning to classify Text Using Support Vector Machines – Methods, Theory, and Algorithms, 2002