International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 7, No. 6 (2014), pp. 75-84
http://dx.doi.org/10.14257/ijsip.2014.7.6.07

A Survey of Automatic Extraction of Personal Name Alias from the Web

A. Muthusamy[1] and A. Subramani[2]

[1]Assistant Professor, Department of Computer science, K.S.R. College of Arts and Science, Tiruchengode, Tamilnadu
[2]Professor & Head, Department of M.C.A, K.S.R. College of Engineering, Tiruchengode, Tamilnadu
muthusamy.arumugam@gmail.com, subramani.appavu@gmail.com

Abstract

The survey paper explains about the extraction and retrieval of personal name alias using various techniques from the web with the help of web crawls. The existing methods help to improve the depth of knowledge relevant to alias extraction and retrieval process. It also describes about how the aliases are ranked, then page counts on the web, word co-occurrence using anchor text and techniques like term frequency (tf), inverse document frequency (idf), log likelihood ratio. Chi-squared tests etc.., are used for measuring the association and similarities between words. The existing method consists of pattern extraction algorithm or string matching algorithm for extracting patterns from snippets instead of using these algorithms. The survey helps to discover a proposed method as graph mining to extract personal name aliases from the web.

Keywords: *Text mining, Information extraction, Web text analysis, Sentiment analysis*

1. Introduction

Now a day's many celebrities like actor, actors, sports men and women, business person etc.., are all on the web. They are using alias name, nickname in order to hide their identity on the web. This survey paper mainly focus on information extraction or information retrieval of an individual person with related attributes like person name place of birth, date of birth, location etc.., by retrieving the information on the web is a difficult process these will return some noisy pattern, in correct alias and so on. Then the strength of the association measure between words is also calculated with the help of statistical methods. Finally Support Vector Machine (SVM) is used to integrate the Mean Reciprocal Rank (MRR) of the system.

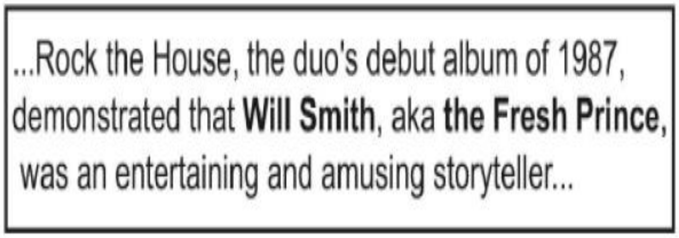
2. Literature Survey

Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka [1] proposed Lexical pattern-based approach to extract aliases of a given name using snippets returned by a web search engine. The lexical patterns are generated automatically using a set of real world name alias data. To select the best aliases among the extracted candidates, we propose numerous ranking scores based upon three approaches: lexical pattern frequency, word co-occurrences in an anchor text graph, and page counts on the web. Moreover, using real-world name alias data, we train a ranking support vector machine to learn the optimal combination of individual ranking scores to construct a robust alias extraction method. We compare the proposed method against numerous baselines and previously proposed name alias extraction methods on three data sets: an English personal names data set, an English place names data set, and a Japanese personal

ISSN: 2005-4254 IJSIP

names data set. Moreover, we evaluate the aliases extracted by the proposed method in an information retrieval task and a relation extraction task. The extracted pattern [1] contains words, symbols and punctuation markers. In web snippets, candidates extracted by lexical pattern might include some invalid aliases. S. Sekine and J. Artiles [2] **Clustering:** Grouping the web pages referring to the same person. **Extraction:** Extracting the attributes for each of the persons sharing the same name. Some of the web people services are zoominfo.com / spock.com / 123people.com. The extracted attributes[2] contains description of the attribute class as DOB, birth of place, other name, occupation, affiliation, award, school, major, degree, mentor, nationality, relatives, phone, fax, e-mail and websites. These are the attributes list retrieved from each cluster of the document (display any one attribute). The attribute information retrieved from the cluster is incomplete. **Example** "director" when it should say "director marketing". The web page is unreadable. If the web page is readable but the specified person is not on this page. If the attribute contains a correct value but it includes irrelevant information. **Example** "CEO in 1982" when it should say "CEO". E. Amigo, J. Gonzalo, J. Artiles, and F. Verdejo [3] **Cluster homogeneity:** it states that the clusters must be homogeneous. **Cluster completeness:** An evaluation metric must hold and refer to the basic goals of a clustering system: keeping items from the same category together and keeping items from different categories apart. **rag bag:** Indeed, for many practical situations it is useful to have a "rag bag" of items which cannot be grouped with other items (think of "miscellaneous", "other", "unclassified" categories); it is then assumed that such a set contains items of diverse genre. Instance the mass error or the number of noise clusters [3] are related simultaneously with the concepts of homogeneity, completeness and Cluster size versus quantity. Therefore, it is not easy to identify the need for satisfying specific constraints in specific clustering applications.

Danushka Bollegala, Yutaka Matsuo, Mitsuru Ishizuka, Taiki Honma [4-5] proposed automatically extracted lexical syntactic patterns from text snippets to select the alias from text snippets pattern extraction algorithm is used. Page count on the web is used to measure co-occurrences of two words. Example: "apple" and "banana" are the two words passed to the web search engine it returns similarly 288,000,000 text snippets. By identifying the exact set of words [4] that convey the semantic relationship between two entities is a difficult problem, it requires deeper semantic analysis. While sending name * alias to the web search engine it retrieves the information from text snippets with some noisy data [5, 40-44, 39].



...Rock the House, the duo's debut album of 1987, demonstrated that **Will Smith**, aka **the Fresh Prince,** was an entertaining and amusing storyteller...

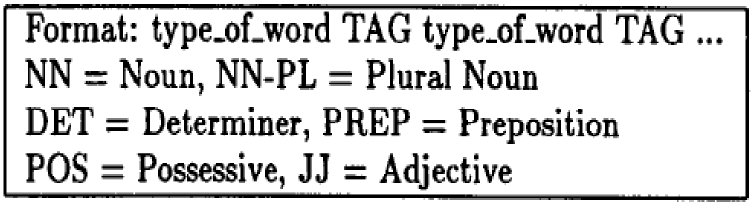
Figure 1. Snippet Returned for the Query Will Smith * the Fresh Prince

C. Galvez and F. Moya-Anegon [6] involved in identification of personal names (PNs) including Natural Language Processing, Information extraction and Information Retrieval (IR) and String-matching algorithms. The Analysis and recognition of the variants is very high through slightly hampered by a problem of over analysis owing to the fact that some strings contain errors. An inherent limitation of such string matching [6] approaches is that they cannot identify aliases. D. Bollegala, Y. Matsuo, and M. Ishizuka [7] the techniques involved in measuring similarities between words are pattern extraction, page count and word co-occurrence. The extracted words [7] might occur arbitrarily *i.e.*, random chance on some pages for those reasons page counts alone are unreliable when measuring semantic similarly. Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka [8] proposed Social networks or Referral web employs several advanced techniques to extract relations of persons, to detect groups of persons, and to obtain keywords for a person. Search engines, especially Google, are used to measure co-occurrence of information and obtain Web documents. Social networks are obtained using analyses of Web pages, e-mail messages, and publications and self-created profiles. POLYPHONET [8] uses co-occurrence information to search the entire web for a person's name. The extract keyword algorithm will collect documents retrieved by a person name and obtain a set of words, phrases as candidates for keywords. If two names co-occur in the same line, they are classified as co-authors [8].

T. Hokama and H. Kitagawa [9] 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". In information extraction and knowledge extraction [9] for a particular object is that the same objects are referred to in different ways in different web documents. P. Mika [10] proposed community based ontology extraction from web pages. del.icio.us is a social bookmarking tool. Much like the similar functions of browsers, del.icio.us allows users to manage a personal collection of links to web sites and describe those links with one or more keywords. Unlike stand-alone tools, del.icio.us is a web based system that allows users to share bookmarks with each other. Bookmarks can be browsed by user, by keywords (tags) or by a combination of both criteria. Further, the user interface encourages exchange by showing how bookmarks are linked together via users and tags. In terms of the Actor-Concept-Instance model, registered users of del.icio.us are the actors who create or remove associations between terms and web pages (instances) by adding or deleting bookmarks. Case sensitivity and the use of punctuation marks further pollute the del.icio.us namespace.

J. Artiles, J. Gonzalo, and F. Verdejo proposed Agglomerative Vector Space clustering algorithm [11], used to evaluate similar tasks], and does not require fixing the number of clusters (K) a priori. The clustering method has been tested using two approaches to build the vector representation of the documents: the first one (full text) uses all the textual contents of the web page as input for the algorithm, and the second one (snippets) only considers the snippets in the ranked lists provided by Google. Words were stemmed using Porter's algorithm [11]. R. Bekkerman and A. McCallum [12] G. Mann and D. Yarowsky [13] proposed personal name disambiguation the goal is to disambiguate various people that share the same name (namesakes) [12, 13]. However, the name disambiguation problem differs fundamentally from that of alias extraction. Because in name disambiguation the objective is to identify the different entities that are referred by the same ambiguous name [12, 13] in alias extraction, the authors are interested in extracting all references to a single entity from the web [12, 13]. In this method many noisy and incorrect aliases are extracted using this pattern. In String matching approaches is that they cannot identify aliases, which share no words or letters with the real name [12, 13]. T. Kudo, K. Yamamoto, and Y. Matsumoto proposed a simple approach would be to let a character be a token (*i.e.*, character-based Begin/Inside tagging) so that boundary ambiguity never occurs. B/I tagging is not a standard method in 20-year history of corpus-based Japanese morphological analysis [14]. This is because B/I tagging cannot directly reflect lexicons which contain prior knowledge about word segmentation. We cannot ignore a lexicon since over 90% accuracy can be achieved even using the longest prefix matching with the lexicon. The lexicon gives a tractable way to build a lattice from an input sentence. A lattice represents all candidate paths or all candidate sequences of tokens, where each token denotes a word with its part of-speech. Begin/Inside tagging [14] produces a number of redundant candidates which makes the decoding speed slower. M. Bilenko and R. Mooney [15] proposed a method to learn a string similarity measure to detect duplicates in bibliography databases. They present two such string similarity measures. The first one utilizes the Expectation-Maximization (EM) algorithm for estimating the parameters of a generative model based on string edit distance with affine gaps. The other string similarity measure 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 measure based on vector-space representation is more appropriate for fields that contain longer strings with more global variations. However, an inherent limitation of such string matching approaches is that they cannot identify aliases. Some previous work has addressed the problem of identifying duplicate records, where it was referred to as record linkage [17, 18], the merge/purge problem [19], duplicate detection [20, 21], hardening soft databases [22], reference matching [24], and entity name clustering and matching [23]. Typically, standard string similarity metrics such as edit distance [26] or vector-space cosine similarity [25] are used to determine whether two values or records are alike enough to be duplicates. Some more recent work [23, 21, 27] has investigated the use of pairing functions that combine multiple standard metrics. T. Joachims [28] proposed clickthrough data in search engines can be thought of as triplets (q, r, c) consisting of the query q, the ranking r presented to the user, and the set c consists of links the user clicked on. Support Vector Machine (SVM) algorithm that leads to a convex program and that can be extended to non-linear ranking functions. While clickthrough data is typically noisy [28] and clicks are not “perfect” relevance judgments, the clicks are likely to convey some information. An approach to learning retrieval functions by analyzing which links the users click on in the presented ranking. This leads to a problem of learning with preference examples like”for query q, document d_a [28] should be ranked higher than document d_b [28]”.

P. Turney proposed the unsupervised learning algorithm [29] has three steps: (1) extract phrases containing adjectives or adverbs, (2) estimate the semantic orientation of each phrase, and (3) classify the review based on the average semantic orientation of the phrases. PMI-IR uses Pointwise Mutual Information (PMI) and Information Retrieval (IR) to measure the similarity of pairs of words or phrases. The se- mantic orientation of a given phrase is calculated by comparing its similarity to a positive reference word (“excellent”) with its similarity to a negative reference word (“poor”).The 410 reviews yielded 10,658 phrases, so the total time required to process the corpus was roughly 106,580 seconds, or about 30 hours [29]. This might appear to be a significant limitation. T. Hisamitsu and Y. Niwa [30] proposed an effective measure of term weighing for word selection from a set of tretrieved documents. This measures uses combinatorial probability and measures like term frequency – inverse document frequency (tf-idf), Log Likelihood Ratio (LLR), chi squared (CS), Hypergeometric distributions (HGS). If the words are appropriately selected the accuracy of tasks such as document similarity calculation in IR or text categorization or clustering can be improved because irrelevant words affecting the accuracy of these tasks are excluded. HGS [30] does not need normalization. M. Berland and E. Charniak [31] proposed the Lexical patterns that tend to indicate part-whole relations. We find possible patterns by taking two words that are in a part whole relation.



Format: type_of_word TAG type_of_word TAG ...
NN = Noun, NN-PL = Plural Noun
DET = Determiner, PREP = Preposition
POS = Possessive, JJ = Adjective

Figure 2. Lexical Pattern to Indicate the Part-whole Relation

The extracted lexical patterns [31] opposed to complete noun phrases this occasionally causes problem. Idiomatic phrases like "a jalopy of a car" or the son of a gun" provide problems that are not easily weeded out. Depending on the data, these phrases can be as prevalent as the legitimate parts. In some cases problems arose because of tagger mistakes. Example: "re-enactment" would be found as part of a "car" using pattern B in the phrase "the re-enactment of the car crash" if "crash" is tagged as a verb. The most persistent problem is sparse data [31], which is the source of most of the noise. C. Manning and H. Schutze [32] proposed Lexical semantics with word meanings is defined by the relations. Common relations are those establishing a hierarchy (Wordnet). Hyperonyms: word with a more general meaning, *e.g.* animal, cat. Hyponym: word with a more specific meaning, *e.g.* cat, animal. Other relations are: Meronyms (part-of): tire, car. Holonyms (has-part): tree, leaf. Antonyms: words with opposite meanings: hot, cold. Synonyms: words with same or similar meanings: car, automobile. Homonyms: two word referring to different unrelated concepts: bank. Polysemous: a word with different related senses: branch. Lexical ambiguity due to homonymy and polisemy. When applying NLP [32] to solving real problems there are a number of low levels processing issues. Given a tag set, part of speech tagging is the problem of assigning the correct grammatical category to a word. Most words may have different POS [32], depending on context of use. Sweet may be noun or adjective, porta may be verb or noun.

M. Mitra, A. Singhal, and C. Buckley [33] proposed this precision-enhancing step into the usual adhoc feedback process, 1) To use K (say 20) documents in the feedback process, retrieve a large number T (say 50) of documents using the original user query. 2) For each retrieved document, compute a new similarity score Sim_{new} based on the occurrence of additional relevance indicators in the document. 3) Rerank the T retrieved documents based on Sim_{new}, breaking ties by the original similarities. 4) Select the top K documents in the new ranking and use them in the Rocchio relevance feedback process to expand the query [33]. Finally, Use the expanded query to retrieve the final list of documents returned to the user. If a large fraction of the documents assumed, then the words added to the query are likely to be unrelated to the topic and the quality of the documents retrieved using the expanded query is likely to be poor [33]. A. Bagga and B. Baldwin [34] proposed a cross-document coreference resolution algorithm by first performing within document coreference resolution for each individual document to extract coreference chains, and then, clustering the coreference chains under a vector space model to identify all mentions of a name in the document set. However, the vastly numerous documents on the web render it impractical to perform within document coreference resolution to each document separately, and then, cluster the documents to find aliases. Ideally, people who work in the same field should be clustered into the same group. We used the B-CUBED metric to evaluate the clustering results. The B-CUBED evaluation metric was originally proposed for evaluating cross-document coreference chains. Many noisy and incorrect aliases are extracted by using this pattern [34]. T. Dunning [35] proposed textual analysis it can be done effectively with very much smaller volumes of text than is necessary for conventional tests based on assumed normal distributions, and it allows comparison to be made between the significance of the occurrences of both rare and common phenomenon. Binomial distributions arise commonly in statistical analysis when the data to be analyzed are derived by counting the number of positive outcomes of repeated identical experiments. In text [35], each comparison is clearly not independent of all others, but the dependency falls off rapidly with distance. F. Smadja [36] proposed a lexicographic tool, Xtract. Expressions are called as collocations. Collocations vary tremendously in the number of words involved, in the syntactic categories of the words, in the syntactic relations between the words, and in how rigidly the individual words are used together. This creates a problem for persons not familiar with the sublanguage as well as for several machine applications such as language generation [36]. In text, each comparison is clearly not independent of all others, but the dependency falls off rapidly with distance. M. Hearst [37] proposed to find new patterns automatically, using the following procedure: 1) Decide on a lexical relation R example: “group / member”. 2) Gather a list of terms for which this relation is known to hold, example: “England-country” this list can be found automatically using the method described here, bootstrapping from patterns or existing lexicon or knowledge base. 3) Find places in the corpus where these expressions occur syntactically near one another and record the environment. 4) Find the commonalities among these environments and hypothesize that common ones yield patterns that indicate the relation of interest. 5) Once a new pattern has been positively identified, use it to gather more instances of the target relation and go to step2. Brent’s algorithm lies in the configuration of interest, where noun phrases are the source of ambiguity, it uses only sentences, which have pronouns in the crucial position, since pronouns do not allow this ambiguity. This approach is quite effective, but the disadvantage is that it isn’t clear that it is applicable to any other tasks [37]. K. Church and P. Hanks [38] proposed statistical description has a large number of potentially important applications, including: (a) constraining the language model both for speech recognition and optical character recognition (OCR), (b) providing disambiguation cues for parsing highly ambiguous syntactic structures such as noun compounds, conjunctions, and prepositional phrases, (c) retrieving texts from large databases (e.g. newspapers, patents), (d) enhancing the productivity of computational linguists in compiling lexicons of lexico syntactic facts, and (e) enhancing the productivity of lexicographers in identifying normal and conventional usage. Concordance analysis is still extremely labor-intensive and prone to errors of omission [38].

3. Related Work

Defining a framework for pattern extraction with the help of web search engines by using graph mining then ranking algorithm using support vector machine is generated in order to extract the lexical patterns, the strength of the association measures between words can be calculated by statistical analysis of the text. We evaluate these methods on two data sets English personal name data set and English place name data set. Then the mean reciprocal rank is calculated which is significantly higher than that of previous ones.

4. Conclusion

The survey paper concludes that the previous methods for extracting personal name alias from the web crawls, social networks. The extracted pattern contains noisy of data, in correct alias, punctuations, markers, typographical errors, misspellings, abbreviations *etc.*, is also represented with examples to overcome this problem graph mining is introduced to extract the pattern in XML format. While using semi structured mining the tags are custom defined but this is not supported in HTML. The next step is the comparative analysis of existing algorithm with proposed one. In future, we need to model a lexico syntactical pattern for extracting the alias from snippets then word co-occurrence for measuring the association between words.

Acknowledgement

I would like to express gratitude to my supervisor Dr. A. Subramani and the management of K.S.R Educational Institutions for the valuable guidance and support.

References

[1] D. Bollegala, Y. Matsuo and M. Ishizuka, “Automatic Discovery of Personal Name Aliases from the Web,” IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 6, (**2011**) June.
[2] 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**).
[3] E. Amig_o, J. Gonzalo, J. Artiles and F. Verdejo, “A comparison of extrinsic clustering evaluation metrics based on formal constraints”, Information Retrieval, (**2008**).
[4] Measuring Semantic Similarity between Words using Web Search Engines Danushka Bollegala, Yutaka Matsuo, Mitsuru Ishizuka, (**2008**) July 22.
[5] Identification of Personal Name Aliases on the Web Danushka Bollegala, Taiki Honma, Yutaka Matsuo, Mitsuru Ishizuka, (**2008**).
[6] C. Galvez and F. Moya-Anegon, “Approximate Personal Name- Matching through Finite-State Graphs,” J. Am. Soc. for Information Science and Technology, vol. 58, pp. 1-17, (**2007**).
[7] D. Bollegala, Y. Matsuo and M. Ishizuka, “Measuring Semantic Similarity between Words Using Web Search Engines,” Proc. Int’l World Wide Web Conf. (WWW ’07), (**2007**), pp. 757-766.
[8] 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 ’06, (**2006**).
[9] T. Hokama and H. Kitagawa, “Extracting Mnemonic Names of People from the Web,” Proc. Ninth Int’l Conf. Asian Digital Libraries (ICADL ’06), (**2006**), pp. 121-130.
[10] P. Mika, “Ontologies Are Us: A Unified Model of Social Networks and Semantics,” Proc. Int’l Semantic Web Conf. (ISWC ’05), (**2005**).
[11] J. Artiles, J. Gonzalo and F. Verdejo, “A Testbed for People Searching Strategies in the WWW,” Proc. SIGIR ’05, (**2005**), pp. 569-570.
[12] R. Bekkerman and A. McCallum, “Disambiguating Web Appearances of People in a Social Network,” Proc. Int’l World Wide Web Conf. (WWW ’05), (**2005**), pp. 463-470.
[13] G. Mann and D. Yarowsky, “Unsupervised Personal Name Disambiguation,” Proc. Conf. Computational Natural Language Learning (CoNLL ’03), (**2003**), pp. 33-40.
[14] 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**).
[15] M. Bilenko and R. Mooney, “Adaptive Duplicate Detection Using Learnable String Similarity Measures,” Proc. SIGKDD ’03, (**2003**).
[16] U. Y. Nahm and R. J. Mooney, “Using information extraction to aid the discovery of prediction rules from texts”, In Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000) Workshop on Text Mining, Boston, MA, (**2000**) August.
[17] H. B. Newcombe, J. M. Kennedy, S. J. Axford and A. P. James, “Automatic linkage of vital records”, Science, vol. 130, (**1959**), pp. 954–959.
[18] W. E. Winkler, “The state of record linkage and current research problems”, Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC, (**1999**).
[19] M. A. Hern´andez and S. J. Stolfo, “The merge/purge problem for large databases”, In Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data (SIGMOD-95), pp. 127–138, San Jose, CA, (**1995**), May.
[20] A. E. Monge and C. P. Elkan, “An efficient domain-independent algorithm for detecting approximately duplicate database records”, In Proceedings of the SIGMOD 1997 Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 23–29, Tuscon, AZ, (**1997**), May.
[21] S. Sarawagi and A. Bhamidipaty, “Interactive deduplication using active learning”, In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), Edmonton, Alberta, (**2002**).
[22] W. W. Cohen, H. Kautz and D. McAllester, “Hardening soft information sources. In Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000), Boston, MA, (**2000**) May.

[23] W. W. Cohen and J. Richman, Learning to match and cluster large high-dimensional data sets for data integration. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), Edmonton, Alberta, (**2002**).

[24] A. K. McCallum, K. Nigam, and L. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching", In Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000), pp. 169–178, Boston, MA, (**2000**) August.

[25] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval", ACM Press, New York, (**1999**).

[26] D. Gusfield., "Algorithms on Strings, Trees and Sequences", Cambridge University Press, New York, (**1997**).

[27] S. Tejada, C. A. Knoblock, and S. Minton, "Learning domain-independent string transformation weights for high accuracy object identification", In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), Edmonton, Alberta, (**2002**).

[28] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. ACM SIGKDD '02, (**2002**).

[29] P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," Proc. Assoc. for Computational Linguistics (ACL '02), (**2002**), pp. 417-424.

[30] T. Hisamitsu and Y. Niwa, "Topic-Word Selection Based on Combinatorial Probability," Proc. Natural Language Processing Pacific-Rim Symp. (NLPRS '01), (**2001**), pp. 289-296.

[31] M. Berland and E. Charniak, "Finding Parts in Very Large Corpora," Proc. Ann. Meeting of the Assoc. for Computational Linguistics (ACL '99), (**1999**), pp. 57-64.

[32] C. Manning and H. Schutze, "Foundations of Statistical Natural Language Processing", MIT Press, (**1999**).

[33] M. Mitra, A. Singhal, and C. Buckley, "Improving Automatic Query Expansion," Proc. SIGIR '98, (**1998**), pp. 206-214.

[34] A. Bagga and B. Baldwin, "Entity-Based Cross-Document Coreferencing Using the Vector Space Model," Proc. Int'l Conf. Computational Linguistics (COLING '98), (**1998**), pp. 79-85.

[35] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," Computational Linguistics, vol. 19, (**1993**), pp. 61-74.

[36] F. Smadja, "Retrieving Collocations from Text: Xtract," Computational Linguistics, vol. 19, no. 1, (**1993**), pp. 143-177.

[37] M. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," Proc. Int'l Conf. Computational Linguistics (COLING '92), (**1992**), pp. 539-545.

[38] K. Church and P. Hanks, "Word Association Norms, Mutual Information and Lexicography," Computational Linguistics, vol. 16, (**1991**), pp. 22-29.

[39] R. C. Bunescu and R. J. Mooney, "Learning to extract relations from the web using minimal supervision", In Proc. of ACL'07, (**2007**), pp. 576-583.

[40] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida and M. Ishizuka, "Polyphonet: An advanced social network extraction system", In Proc. of WWW'06, (**2006**).

[41] J. Mori, Y. Matsuo, and M. Ishizuka, "Extracting keyphrases to represent relations in social networks from the web", In Proc. of IJCAI'07, (**2007**), pp. 2820-2852.

[42] M. Fleischman and E. Hovy, "Multi-document person name resolution", In Proc. of 42nd ACL, Reference Resolution Workshop, (**2004**).

[43] B. Aleman-Meza, M. Nagarajan, and I. Arpinar, "Ontology-driven automatic entity disambiguation in unstructured text", In Proc. of ISWC'06, (**2006**).

[44] T. Pedersen, A. Kulkarni, R. Angheluta, Z. Kozareva, and T. Solorio, "An unsupervised language independent method of name discrimination using second order co-occurance features", In Proc. of CICLing'06, (**2006**).

Authors

A. Muthusamy, he is currently working as a Assistant Professor, Department of Computer science, K.S.R. College of Arts and Science, Tiruchengode and as a Research Scholar in Bharathiar University, Coimbatore. He received his MCA degree from Anna University, Chennai. He published 1 technical paper in National Conference, 3 National / International Journal. His area of research includes Data Mining, Web Mining.

A. Subramani, he is currently working as a Professor and Head, Department of Computer Applications, K.S.R. College of Engineering, Tiruchengode and as a Research Guide in various Universities. He received his Ph.D. Degree in Computer Applications from Anna University, Chennai. He is a Reviewer of 10 National / International Journals. He is in the editorial board of 6 International / National Journals. He is an Associate Editor of Journal of Computer Applications. He has published more than 30 technical papers at various International, National Journals and Conference proceedings. His areas of research includes High Speed Networks, Routing Algorithm, Soft computing, Wireless Communications, Mobile Ad-hoc Networks.