A Lexical Pattern Based Method for Searching of Personal Name

K.S.Lakshmi[*], Kumar.Vasantha[#], Prof. C.MohanRao[$]
[*]*M.Tech Scholar,*[#] *HOD,*[$] *Professor and Principal ,*
[*#$]*Dept of CSE, Avanthi Institute of Engineering and Technology*

ABSTRACT:Searching of names on the web is typical task because that data is refers more alias names in the web. We have to retrieve correct information about given individual name. So we have to find out the similarities between the names, such that we can find the appropriate information about the individual name. We introduced lexical pattern based approach, first it find lexical patterns and co-occurrences based on the page counts and point wise mutual information between two candidates of lexical patterns and find the probability such mean reciprocal rank(MRR) based on this , the highest MRR candidate have the correct alias name for the given name to search.

INTRODUCTION:

Efficient and accurate identification of aliases of a given name of the person is useful in various web related tasks such as retrieval of information sentiment analysis and personal name disambiguation, and relation extraction. Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources and Searches can be based on metadata or on full-text (or other content-based) indexing Automated information retrieval systems are used to reduce what has been called "information overload" approach when Many universities and public libraries use IR systems to provide books access, published journals and other documents as follows, Web search engines are the most visible IR applications and An information retrieval process begins when a user enters a query into the system and Queries are formal statements of needs of information for example search strings in web search engines[1][2]. In our information retrieval process a query does not uniquely identify a single object in the collection. Instead, several objects may compares the query process, may be with different degrees of relevancy and An object is an entity that is represented by information in a database and User queries are matched against the information of database. Depends on the application the data objects may be, consider an example, text documents, images, audio, mind maps or videos or other. Very Often the documents themselves are not kept or stored directly in the IR system but are instead represented in the system by document surrogates or metadata[3].

Sentiment Analysis is a Natural Language Processing and Information Extraction task that aims to obtain writer's feelings expressed in positive or negative comments requests and questions, by analyzing a large numbers of documents. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall tonality of a document. Now a days ,exponential increase in the Internet usage and exchange of public opinion is the driving force behind Sentiment Analysis today. The Web is a huge repository of structured and unstructured data. The analysis of this data to extract latent public opinion and sentiment is a challenging task. The analysis of sentiments may be document based where the sentiment in the entire document is summarized as negative , positive or objective and It can be sentence based where individual sentences bearing sentiments and in the text are classified. SA can be phrase based where the phrases in a sentence are classified according to polarity[4][5]. Sentiment Analysis identifies the phrases in a text that bears some sort of sentiment and the author may speak about some objective facts or subjective opinions. It is necessary to distinguish between the two. SA finds the subject towards whom the directed sentiment. A text may contain many entities but it is necessary to find the entity towards which the directed sentiment and It identifies the polarity and degree of the sentiment. Sentiments are classified as objective (facts), positive (denotes a state of happiness and bliss or satisfaction on part of the writer) or negative (denotes a state of sorrow dejection or disappointment on part of the writer), by the sentiments can further be given a score based on their degree of positivity, negativity or objectivity[6][7].

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 "Jim Clark" may potentially refer to hundreds or thousands of distinct individuals. Each different referent typically has some distinct contextual characteristics. And these characteristics can help distinguish and resolve the referents when the surface names appear in online documents.

A relationship extraction task requires the detection and classification of semantic relationship mentions within a set of artifacts and typically from text or XML documents and its task is very similar to that of information extraction (IE) but IE additionally requires the removal of repeated relations (disambiguation)

and generally refers to the extraction of many different relationships[12].

Early IR systems were boolean systems which allowed users to specify their information need using a complex combination of boolean ANDs, ORs and NOTs[13]. Boolean systems have several shortcomings e.g., and there is no inherent notion of document ranking, and it is very hard for a user to form a good search request and Even though boolean systems usually return matching documents in some order, e.g., ordered by date, or some other document feature, relevance ranking is often not critical in a boolean system. Even though it has been shown by the research community that boolean systems are less effective than ranked retrieval systems, many power users still use boolean systems as they feel more in control of the retrieval process. However, most everyday users of IR systems expect IR systems to do ranked retrieval. IR systems rank documents by their estimation of the usefulness of a document for a user query. Most IR systems assign a numeric score to every document and rank documents by this score. Several models have been proposed for this process. The three most used models in IR research are the vector space model and the probabilistic models, and the inference network model[14][15].

RELATED WORK

A) Lexical patterns

We use two different text pre-processing methods which automatically assign linguistic information tosentences. The first pre-processing method has the advantage of offering a fast analysis of the data but its results are less elaborate than those of the second method. The first method consists of three steps:

• **Segmentation :** sentence boundaries are detected and punctuation signs are separated from words

• **Tagging of part of speech** : part-of-speech classes like noun and verb are assigned to words

• **Lemmatization:** words are reduced to their basic form (lemma) Like in the work of Snow et al. (2005), the target phrases for hyponym extraction are two noun phrases, with a maximum of three tokens in between and one or two optional extra tokens (a nonhead token of the first noun phrase and/or one of the second noun phrase).

The lexical pre-processing method uses two basic regular expressions for finding noun phrases: Determiner? Adjective* Noun+ and Proper Noun+. It assumes that the final token of the matched phrase is the head. Here is one set of four patterns which can be derived from the example sentence:

1. NP in NP
2. large NP in NP
3. NP in north NP
4. large NP in north NP

The patterns contain the lemmas rather than the words of the sentence in order to allow for general patterns and for the same reason, the noun phrases have been replaced by the token NP and Each of the four patterns will be used as evidence for a possible hyponymy relation between the two noun phrase heads city and England As a novel extension to the work of Snow and we included two additional variants of each pattern in which either the first NP or the second NP was replaced by its head:

5. city in NP
6. NP in England

This enabled us to identify among others appositions as patterns: president NP. Dependency patterns A dependency analysis contains the same three steps used for finding lexical patterns: segmentation, part of- speech tagging and lemmatization. Additionally, it includes a fourth step:

• **Dependency parsing:** find the syntactic dependency relations between the words in each sentence The syntactic analysis is head-based which means that for each word in the sentence it finds another word that dominates it. Here is a possible analysis of the previous example sentence:

Each line contains a lemma, its part-of-speech tag, the relation between the word and its head, the part of-speech tag of its head and the lemma of the head word. Our work with dependency patterns closely follows the work of Snow et al. (2005). Patterns are defined as dependency paths with at most three intermediate nodes between the two focus nouns. Additional satellite nodes can be present next to the two nouns. The dependency patterns contain more information than the lexical patterns and here is one of the patterns that can be derived for the two noun phrases large cities and northern England in the example sentence:

The pattern defines a path from the head lemma city via in, to England. Note that lemma information linking outside this pattern (be at the end of the first line) has been removed and that lemma information from the target noun phrases has been replaced by the name of the noun phrase (NP1 at the end of the second line). For each dependency pattern, we build six variants similar to the six variants of the lexical patterns: four with additional information from the two noun phrases and two more with head information of one of the two target NPs.

Both pre-processing methods can identify phrases like N such as N1 , N2 and N3 as well. Such phrases produce evidence for each of the pairs (N,N1), (N,N2) and (N,N3). These three noun pairs will be included in the data collected for the patterns that can be derived from the phrase. We expect that an important advantage of using dependency patterns over lexical patterns will be that the former offer a wider coverage. In the example sentence, no lexical pattern will associate city with Liverpool because there are too many words in between. However, a dependency pattern will create a link between these two words, via the word as. This will enable the dependency patterns to find out that city is a hyponym of Liverpool, where the lexical patterns are not able to do this based on

the available information. The two pre-processing methods generate a large number of noun pairs associated by patterns.

Like Snow et al. (2005), we keep only noun pairs which are associated by at least five different patterns. The same constraint is enforced on the extraction patterns: we keep only the patterns which are associated by at least five different noun pairs. The data is converted to binary feature vectors representing noun pairs. These are training data for a Bayesian Logistic Regression system, BBR train (Genkin et al., 2004). We use the default settings of the learning system and test its prediction capability in a binary classification task: whether two nouns are related according to hyponymy or not. Evaluation is performed by 10-fold cross validation.

B)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 .

C)The Vector Space Model

The vector space model used for disambiguating entities across documents is the standard vector space model used widely in information retrieval (Salton 89). In this model, each summary extracted by the Sentence Extractor module is stored as a vector of terms. The terms in the vector are in their morphological root form and are filtered for stop-words (words that have no information content like a, the, of, an, ...). If s_1 and s_2 are the vectors for the two summaries extracted from documents D1 and D2, then their similarity is computed as:

$Sim(s_1,s_2) = \sum_{\text{common terms tj}} (w_{1i}-w_{2i})$

wheret_j is a term present in both s_1 and s_2, w_{li} is the weight of the term tj in S1 and w_{1j} is the weight of t_j in s_2.The weight of a term t_i in the vector S_i for a summary is given by

$w_{ij} = tf*log(N/df)/\sqrt{s_{i1}^2 + s_{i2}^2 + ...+s_{in}^2}$

where t_f is the frequency of the term t_i in the sum mary, N is the total number of documents in the collection being examined, and d_f is the number of documents in the collection that the term t_i occurs in is the cosine normalization factor and is equal to the Euclidean length of the vector $S_i.\sqrt{s_{i1}^2 + s_{i2}^2 + ...+s_{in}^2}$The VSM-Disambiguate module, for each summary Si, computes the similarity of that summary with each of he other summaries. If the similarity computed is above a pre-defined threshold, then the entity of interest in the two summaries are considered to be co-referent.

D)Co-Occurrences in Anchor Texts

Anchor texts have been studied extensively in information retrieval and have been used in various tasks such as extraction of synonym and query translation in cross-language information retrieval and ranking and classification of web pages [16]. Anchor texts are particularly attractive because they not only contain concise texts but also provide links that can be considered as expressing a citation and we revisit anchor texts to measure the association between a name and its aliases on the web and Anchor texts pointing to a url provide useful semantic clues related to the resource represented by the url. For example, if the majority of inbound anchor texts of a url contain a personal name and it is likely that the remainder of the inbound anchor texts contain information about aliases of the name, there we use the term inbound anchor texts to refer the set of anchor texts pointing to the same urland We define a name p and a candidate alias x as co-occurring, if x and p appear in two different inbound anchor texts of a url u. Moreover, we define co-occurrence frequency (CF) as the number of different urls in which they co-occur. It is noteworthy that we do not consider co-occurrences of an alias and a name in the same anchor text.

A document consists of sentences of different lengths and if a term appears in a long sentence and it is likely to co-occur with many terms; if a term appears in a short sentence, it is less likely to co-occur with other terms. We consider the length of each sentence and revise our definitions and We denote

- p_g as (the sum of the total number of terms in sentences where g appears) divided by (the total number of terms in the document),
- n_w as the total number of terms in sentences where w appears.

Table 3: Terms with high χ^2

Rank	$\chi 2$	Term	Frequency
1	593.7	Digital	31
2	179.3	computer	16
3	163.1	Imitation	4
4	161.1	game	44
5	152.8	Future	3
6	143.5	Question	39
7	142.8	Internal	15
.	.	Answer	.
.	.	Input signal	.
.	.	.	.
559	0.6	.	2

559 560	0.3 0.1	. Mr. Scan Worse	2 2

Again $n_w p_g$ represents the expected frequency of co-occurrence. However, its value becomes more sophisticated. A term co-occurring with a particular term g∈G has a high $\chi 2$ value. However, these terms are sometimes adjuncts of term g and not important terms. For example, in Table 3, a term "future" or "internal" co-occurs selectively with the frequent term "state," because these terms are used in the form of "future state" and "internal state." Though χ 2 values for these terms are high, "future" and "internal" themselves are not important. Assuming the "state" is not a frequent term, χ 2 values of these terms diminish rapidly. We use the following function to measure robustness of bias values; it subtracts the maximal term from theχ 2 value .

$$X^{t2}(w)=x^2(W)-max_{g\epsilon G}(freq(w,g)-n_wP_g)^2/n_w/n_wP_g$$

Proposed work:

The semantic web is intended to solve the entity disambiguation problem by providing a mechanism to add semantic metadata for entities. However, an issue that the semantic web currently faces is that insufficient semantically annotated web contents are available. Automatic extraction of metadata can accelerate the process of semantic annotation. For named entities, automatically extracted aliases can serve as a useful source of metadata, 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. propose a social network extraction algorithm in which they compute the strength of the relation between two individuals A and B by the web hits for the conjunctive query, "A" and "B". However, both persons A and B 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.

- The semantic web is intended to solve the entity disambiguation problem by providing a mechanism to add semantic metadata for entities.
- Automatic extraction of metadata can accelerate the process of semantic annotation.
- **Extracting Lexical Patterns from Snippets**

Many modern search engines provide a brief text snippet for each search result by selecting the text that appears in the web page in the proximity of the query. Such snippets provide valuable information related to the local context of the query. For names and aliases, snippets convey useful semantic clues that can be used to extract lexical patterns that are frequently used to express aliases of a name

- **Ranking of Candidates**

Considering the noise in web snippets, candidates extracted by the shallow lexical patterns might include some invalid aliases. From among these candidates, we must identify those, which are most likely to be correct aliases of a given name. We model this problem of alias recognition as one of

ranking candidates with respect to a given name such that the candidates, who are most likely to be correct aliases are assigned a higher rank

- **Lexical Pattern Frequency**

  We presented an algorithm to extract numerous lexical patterns that are used to describe aliases of a personal name. As we will see later in Section 4, the proposed pattern extraction algorithm can extract a large number of lexical patterns. If the personal name under consideration and a candidate alias occur in many lexical patterns, then it can be considered as a good alias for the personal name. Consequently, we rank a set of candidate aliases in the descending order of the number of different lexical patterns in which they appear with a name. The lexical pattern frequency of an alias is analogous to the document frequency (DF) popularly used in information retrieval.

- **Co-Occurrences in Anchor Texts**

  Anchor texts have been studied extensively in information retrieval and have been used in various tasks such as synonym extraction, query translation in cross-language information retrieval, and ranking and classification of web pages . Anchor texts are particularly attractive because they not only contain concise texts, but also provide links that can be considered as expressing a citation. We revisit anchor texts to measure the association between a name and its aliases on the web. Anchor texts pointing to a url provide useful semantic clues related to the resource represented by the url. For example, if the majority of inbound anchor texts of a url contain a personal name, it is likely that the remainder of the inbound anchor texts contain information about aliases of the name. Here, we use the term inbound anchor texts to refer the set of anchor texts pointing to the same url.

A) Lexical patterns extraction

In our approach we introduced lexical pattern based approach. First the given input id divided into tokens. Those tokens are so called as lexical patterns.

Algorithm 1: Extract Patterns(S)

Comment: S is a set of (NAME, ALIAS) pairs
P←*null*
For each (NAME,ALIAS) ϵ S
Do
    D← Get Snippets (“NAME *ALIAS”)
    for each snippet d ϵD
    do*P*←*P* + Create Pattern(d)
return (*P*)

Algorithm 2: Extract Candidates(NAME,P)

Comment: *P is the set of patterns*
C←*null*
For each patterns p ϵ *P*
Do
    D← Get Snippets (“NAME *ALIAS”)
    for each snippet d ϵ D
    do*C*←*C* + Get N grams*(d,NAME,P)*
return (*C*)

B) Ranking the candidates

After extracting lexical patterns, we have to check alias names of the original names correct or not. Then we define ranking scores of the each and every lexical pattern. For this we firstly define the co-occurrences of the given names and the alias names. Co-occurrence means that two patterns have some relation for example 'student' and 'class room'. There is a relation between student and classroom for that we can follow similarity measures.

a)Coefficients for Measuring Association

The following are a few of the many measures of association used with chi-square and other contingency analyses of table . While using the chi-square statistic and these coefficients can be helpful in interpreting the relationship between two variables once statistical significance has been established and the logic for using measures of association is as follows:

Even though a chi-square test may show statistical significance between two variables and the relationship between those variables may not be substantively important. These and many other measures of association are available to help evaluate the relative strength of a statistically significant relationshipand In most cases, they are not used in interpreting the data unless the chi-square statistic first shows there is statistical significance (i.e., it doesn't make sense to say there is a strong relationship between two variables when your statistical test shows this relationship is not statistically significant).

a.Nominal and Ordinal Variables

1)Phi

Only used on 2x2 tables of contingency and Interpreted as a measure of the relative (strength) of an association between two variables lies from 0 to 1.

$$Phi = \sqrt{\frac{x^2}{n}}$$

2)Pearson's Contingency Coefficient (C)

It is interpreted as a measure of the relative (strength) of an association between two variables and the coefficient will always be less than 1 and varies according to the number of rows and columns.

$$C = \sqrt{\frac{x^2}{(n + x^2)}}$$

3)Cramer's V Coefficient (V)

Useful for comparing multiple X^2 test statistics and is generalizable across contingency tables of varying sizes. It is not affected by sample size and therefore is very useful in situations where you suspect a statistically significant chi-square was the result of large sample size instead of any substantive relationship between the variables. It is interpreted as a measure of the relative (strength) of an association between two variables. The coefficient ranges from 0 to 1 (perfect association) and other In practice, you may find that a Cramer's V of .10 provides a good minimum threshold for suggesting there is a substantive relationship between two variables.

$$V = \sqrt{X^2/n\,(q-1)}$$

Whereq = smaller # of rows or columns

Describing Strength of Association

Characterizations	
>.5	high association
.3 to .5	moderate association
.1 to .3	low association
0 to .1	little if any association

4)Proportional Reduction of Error (PRE)

Lambda

This is a proportional reduction in error (PRE) measure that ranges from 0 to 1 and Lambda indicates the extent to which the independent variable reduces the error associated with predicting the value of a dependent variable. Multiplied by 100 and it represents the percent reduction in error.

5)Ordinal Variables Only Gamma

Another PRE measure ranging from -1 to 1 that estimates the extent errors are reduced in predicting the order of paired cases. Gamma ignores ties.

Kendall's Tau b

Similar to Gamma but includes ties. Ranges from -1 to 1 but since standardization is different from Gamma, it provides no clear explanation of PRE.

6) Inter-rater Agreement

Cohen's Kappa

Measures agreement beyond chance Although a negative value is possible and it commonly ranges from 0 to 1 (perfect agreement) and this measure requires a balanced table where the number of rows is the same as the number of columns. The diagonal cells represent agreement.

A frequently observed phenomenon related to the web is that many pages with diverse topics link to so-called hubs such as Google, Yahoo, or MSN. Two anchor texts might link to a hub for entirely different reasons. Therefore, co-occurrences coming from hubs are prone to noise. To overcome the adverse effects of a hub h when computing co-occurrence measures and we multiply the number of co-occurrences of words linked to h by a factor α(h, p), where α(h,p)=t/d.

Here, t is the number of inbound anchor texts of h that contain the real name p and d is the total number of inbound anchor texts of h. If many anchor texts that link to h contain p (i.e. larger t value), then the reliability of h as a

source of information about p increases and On the other hand, if h has many inbound links (i.e. larger d value), then it is likely to be a noisy hub and gets discounted when multiplied by α(<< 1). Intuitively, in the above expression boosts hubs that are likely to contain information related to p, while penalizing those that contain various other topics.

Based on the web pages we have to find we have to find the measures such as web dice, and point mutual information based on the page counts.

C)Web dice

we compute web dice between a name p and the alias x is given below:

$$WebDice(p,x) = \frac{(2 \times hits(\text{p AND x}))}{(hits(p) + hits(x))}$$

D)Web point mutual information

We compute point mutual information based on the page counts. It is given below:

$$WebPMI(p,x) = \log_2 \frac{(L \times hits(\text{p AND x}))}{(hits(p) + hits(x))}$$

Mean reciprocal rank (MRR) and AP [26] is used to evaluate the different approaches. MRR is defined as follows:

$$MRR = \frac{1}{n}\sum_{i=1}^{n} 1/R_i$$

Therein: Ri is the rank assigned to a correct alias and n is the total number of aliases. AP is defined as follows:

$$AP = \frac{(\sum_{r=1}^{k} Pre(r) \times Rel(r))}{(Noofcorrectaliases)}$$

After calculating all parameters for the given data and the highest values of MRR and AP ranking candidate is the correct alias name for the name.

Experimental Analysis:

Input dataset for alias names searching.

Original Name	Alias Name
sachin	master
sachin	master blaster
sachin	cricket god
sachin	sachin tendulkar
sachin	tendulkar
chiranjeevi	chiru
chiranjeevi	megastar
chiranjeevi	annaya
chiranjeevi	supreem hero
chiranjeevi	mastar
amithabachan	amithab
amithabachan	bigb
amithabachan	badsha

After pattern extraction based on alias name

Search for Original name=sachin and alias name=master

File Name:
index1.html----------> Snippet=is know as
index10.html---------->
index2.html---------->
index3.html---------->
index4.html----------> Snippet=is
index5.html---------->
index6.html---------->
index7.html---------->
index8.html---------->
index9.html---------->

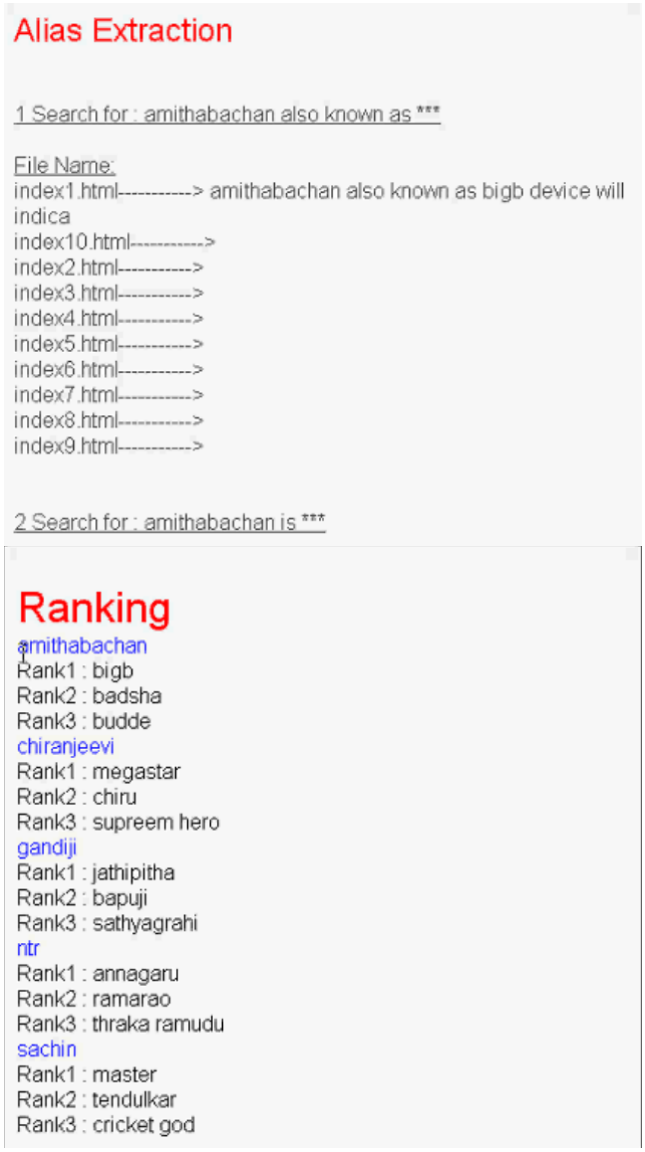
Search for Original name=sachin and alias name=master blaster

File Name:
index1.html---------->
index10.html---------->
index2.html---------->
index3.html---------->
index4.html---------->
index5.html---------->
index6.html---------->
index7.html---------->
index8.html---------->
index9.html---------->

Found snippets are

Snippets
also known as
is
is also known as
is famous as
is know as
is named

Alias names extraction



CONCLUSION:

We introduced lexical patterns based approach for the searching of alias names on the web. We use popular measures for finding the similarities between the patterns. It will give you the high probability of the given candidate pattern. It will give best results of best search and also efficient and simple. It will work on normal English names, the extracted aliases significantly improved recall in a relation detection task. Our method achieves best results after applying mean reciprocal rank on the candidate.

REFERENCES:

[1] R. Guha and A. Garg, "Disambiguating People in Search,"technical report, Stanford Univ., 2004.

[2] J. Artiles, J. Gonzalo, and F. Verdejo, "A Testbed for PeopleSearching Strategies in the WWW," Proc. SIGIR '05, pp. 569-570,2005.

[3] G. Mann and D. Yarowsky, "Unsupervised Personal NameDisambiguation," Proc. Conf. Computational Natural LanguageLearning (CoNLL '03), pp. 33-40, 2003.

[4] R. Bekkerman and A. McCallum, "Disambiguating Web Appearancesof People in a Social Network," Proc. Int'l World Wide WebConf.(WWW '05), pp. 463-470, 2005.

[5] G. Salton and M. McGill, Introduction to Modern InformationRetreival. McGraw-Hill Inc., 1986.

[6] M. Mitra, A. Singhal, and C. Buckley, "Improving AutomaticQuery Expansion," Proc. SIGIR '98, pp. 206-214, 1998.

[7] P. Cimano, S. Handschuh, and S. Staab, "Towards the Self-Annotating Web," Proc. Int'l World Wide Web Conf. (WWW '04),2004.

[8] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H.Takeda, K. Hasida, and M. Ishizuka, "Polyphonet: An AdvancedSocial Network Extraction System," Proc. WWW '06, 2006.

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

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

[11] C. Galvez and F. Moya-Anegon, "Approximate Personal Name-Matching through Finite-State Graphs," J. Am. Soc. for InformationScience and Technology, vol. 58, pp. 1-17, 2007.

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

[13] T. Hokama and H. Kitagawa, "Extracting Mnemonic Names ofPeople from the Web," Proc. Ninth Int'l Conf. Asian Digital Libraries(ICADL '06), pp. 121-130, 2006.

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

[15] M. Berland and E. Charniak, "Finding Parts in Very LargeCorpora," Proc. Ann. Meeting of the Assoc. for ComputationalLinguistics (ACL '99), pp. 57-64, 1999.

[16] S. Chakrabarti, Mining the Web: Discovering Knowledge fromHypertext Data. Morgan Kaufmann, 2003.

[17] G. Salton and C. Buckley, "Term-Weighting Approaches inAutomatic Text Retrieval," Information Processing and Management,vol. 24, pp. 513-523, 1988.

[18] C. Manning and H. Schutze, Foundations of Statistical Natural
Language Processing.MIT Press, 1999.

[19] T. Dunning, "Accurate Methods for the Statistics of Surprise andCoincidence," Computational Linguistics, vol. 19, pp. 61-74, 1993.

[20] K. Church and P. Hanks, "Word Association Norms, MutualInformation and Lexicography," Computational Linguistics, vol. 16,pp. 22-29, 1991.
[21] T. Hisamitsu and Y. Niwa, "Topic-Word Selection Based onCombinatorial Probability," Proc. Natural Language ProcessingPacific-Rim Symp.(NLPRS '01), pp. 289-296, 2001.
[22] F. Smadja, "Retrieving Collocations from Text: Xtract," ComputationalLinguistics, vol. 19, no. 1, pp. 143-177, 1993.
[23] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring SemanticSimilarity between Words Using Web Search Engines," Proc. Int'l
World Wide Web Conf. (WWW '07), pp. 757-766, 2007.
[24] T. Joachims, "Optimizing Search Engines Using ClickthroughData," Proc. ACM SIGKDD '02, 2002.
[25] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying ConditionalRandom Fields to Japanese Morphological Analysis," Proc.
Conf. Empirical Methods in Natural Language (EMNLP '04), 2004.[26] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.ACM Press/Addison-Wesley, 1999.
[27] P. Mika, "Ontologies Are Us: A Unified Model of Social Networksand Semantics," Proc. Int'l Semantic Web Conf. (ISWC '05), 2005.
[28] S. Sekine and J. Artiles, "Weps 2 Evaluation Campaign: Overviewof the Web People Search Attribute Extraction Task," Proc. SecondWeb People Search Evaluation Workshop (WePS '09) at 18th Int'lWorld Wide Web Conf., 2009.

BIOGRAPHIES

Author 1:**K.S.Lakshmi** received B.Tech degree in CSIT from LBRCE, Mylavaram. She Pursuing M.Tech2nd year inComputer Science and Engineering fromAvanthi Institute of Engineering and Technology. Herareas of interests are DBMS, DMDW, Web Mining.

Author 2:**Mr. kumarvasantha,M.Tech(CSE).**
He has obtained M.Tech in Computer Science and Engineering from Jawaharlal Nehru Technological University, Kakinada. He has published 10 papers in National and International journals, His areas of interestsare Data Warehousing, Computer Networks, RDBMS,Web Technologies.

Author 3:**Dr. C. Mohan Rao. M.Tech(CSE),Ph.D.**
He has obtained M.Tech in Computer Science and Technology from Andhra University College of Engineering and awarded Ph.D by Andhra University during 2000. He has 18 years of teaching and research experience and guided number of M.Tech students for their projects. He has published 23 papers in National and International journals. He is guiding 2 research scholars for Ph.D. He received Best Teacher award from JNTU, Kakinada during 2009.

