Sense Based Organization of Descriptive Data

M. Shahriar Hossain, Monika Akbar, and Rafal A. Angryk

Abstract— **In this paper we propose a new technique allowing to map descriptive data into relative distance space, which is based primarily on senses of the terms stored in our data. We use WordNet ontology to retrieve multiple senses of words with the aim of multidimensional representation of data. The focus of this work is mainly on the slicing of available ontology into multiple dimensions where each dimension reflects approximation of a single general sense reflecting broad context of terms/words stored in our document repository. We have concentrated on discovery of appropriate similarity measurements and constructions of data driven dimension. It benefits quality of generated dimensions and provides a clear view of the whole data repository in low dimensional context driven space.**

I. Introduction

HIGH dimensionality of data limits the choice of data mining techniques in many applications. Complex data analysis and mining on huge amounts of data can take a long time, making data analysis impractical and infeasible [26]. This is the reason why different dimensionality reduction techniques are developed so that data mining tasks become more convenient, fast and understandable to human being.

Typically, a large number of words exist in even a moderately sized set of documents resulting in high dimensional text repository [27]. As a result, many mining applications for text data generate impractical and infeasible results. In this paper, we propose a technique to generate sense-based dimensions reflecting broad context of words in a document repository. In our approach, we form a dynamic method which is document set-based. Our aim is to construct the dimensions depending on the terms/keywords in the set of documents. We utilized WordNet [1] ontology as background knowledge to retrieve senses of terms/keywords.

The goal of this work is to utilize ontology as a sense based representation mechanism in multiple dimensions so that the representation of linguistic senses becomes apparent in text repository. Besides, we focus on similarity measurements between words and synstets. A synset is a set of synonyms of a word which provides a broader meaning in sense domain. There are different methods to find out the similarity between two words or synsets. Some similarity measurements are corpus dependent while others are corpus independent. We have chosen WordNet as the background knowledge to retrieve corpus independent similarity measures because it is a lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory [2].

The whole paper is organized in several sections. Section II contains basic literature on WordNet and describes different proximity measures between words and synsets. Besides, it explains Relative Distance Plane (RDP) and Silhouette Coefficient. We propose the basic dimension retrieval algorithm in section III. As measures of proximity are important for our work, we portray some analysis on proximity measures in section IV. We present the experimental results with the dimension retrieval algorithm in section V using a small document set containing 5 text documents and conclude in section VI. Finally, the appendix contains a simple example clarifying the proposed dimension retrieval algorithm.

II. Background and Related Works

In this section we describe WordNet ontology and different measures of words' proximity. The reason why we focus on proximity measures is that similarity measurement is the core of our dimension retrieval technique. Later we show that the change in proximity measure can significantly change the organization of descriptive data.

WordNet divides its whole lexical reference system into five categories: nouns, verbs, adjectives, adverbs, and function words [1]. Function words are basically non-content words like prepositions, conjunctions, etc. that may mislead language processing tasks as they are non-informative. In our work, we have concentrated on nouns, their senses, synsets and coordinate terms only to present our approach in a straightforward manner. In WordNet, synsets are usually connected to other synsets via a number of semantic relations. These relations vary based on the type of word. For example, nouns have five kinds of relations, which are stated below:
(1) *hypernyms*: Y is a hypernym of X if every X is a kind of Y,
(2) *hyponyms*: Y is a hyponym of X if every Y is a kind of X,
(3) *coordinate terms*: Y is a coordinate term of X if X and Y share a hypernym,
(4) *holonym*: Y is a holonym of X if X is a part of Y,
(5) *meronym*: Y is a meronym of X if Y is a part of X.

In WordNet, only adjectives and adverbs are organized as N-dimensional hyperspaces, but nouns and verbs are organized in lexical memory as hierarchies. We shall take the advantage of these hierarchies to retrieve sense and organize concepts in sense based multi-dimensional space.

Sense retrieval can be discussed in the context of Word

M. S. Hossain is a graduate student of Department of Computer Science, Montana State University, Bozeman, MT 59717, USA. (phone: 1-406-209-7103; fax: 1-406-994-4376; e-mail: mshossain@cs.montana.edu).

M. Akbar is also a graduate student of the same department. (e-mail: monika@cs.montana.edu).

R. A. Angryk is a faculty with the Department of Computer Science, Montana State University, Bozeman, MT 59717, USA. (e-mail: angryk@cs.montana.edu).

Sense Disambiguation (WSD) [3]. In computational linguistics, WSD is a problem of determining which sense of a word is used in a given sentence, having a number of distinct senses to choose from. For example, consider the word *bass*, two distinct senses of which are: (1) a type of fish and (2) tones of low frequency. Now take under consideration two sentences: "*The bass part of the song is very moving*", and "*I went fishing for some sea bass*". To a human it is obvious that the first sentence is using the word *bass* in sense 2 above, and in the second sentence it is being used in sense 1. Although this seems obvious to a human, developing algorithms to replicate this human ability is a difficult task. Some interesting works on WSD and mechanism to disambiguate senses from context have been already published. Stevenson et al. [4] describe solution of WSD problem with the interaction of knowledge sources. Their work attempts improvement of disambiguation by interacting several knowledge sources when implementing a sense tagger. The system moves to alternative knowledge source if the sense of a word is not retrieved from one source with satisfactory confidence. The authors report about the accuracy exceeding 94% on their evaluation corpus, which shows that the approach is robust. The approach may however need significant number of knowledge sources, which from our perspective, has been somehow overwhelming. As we have chosen to use only WordNet ontology, we wanted disambiguation process using the same knowledge archive rather than using several knowledge sources.

Montoyo et al. [5] present such a method for disambiguation of nouns in English texts that uses the notion of Specification Marks and employ the noun taxonomy of the WordNet lexical knowledge base. The method resolves the lexical ambiguity of nouns in any sort of text. It relies only on the semantic relations (*hypernymy* and *hyponymy*) and the hierarchic organization of WordNet. The method does not, however, require any sort of training process, no hand-coding of lexical entries, nor the hand-tagging of texts. The intuition underlying this approach is that the more similar two words are, the more informative the most specific concept that subsumes them both, will be. In other words, it uses their lowest common upper bound in the taxonomy.

A. Measurements of Proximity

Sense based representation provides a foundation for words/synsets clustering. Following common clustering principle, we can say that maximizing the intraclass similarity and minimizing the interclass similarity, is the way of proper words' clustering. Clusters are defined as collections of objects whose intraclass similarity is high and interclass similarity is low [6]. As in our approach, metric distance has no use, what we want to use is a sequence of terms in their common dimension and the relative distance between concepts, based on some kind of measures of terms' proximity. In practice, the measurement of proximity between data points is strongly domain-dependent [7]. Yager [7] describes a fundamental distinction between the nearest neighbor cluster distance measure, *Min*, and the furthest neighbor measure, *Max* where the first favors the merging of large clusters while the latter favors the merging of smaller clusters. However, whenever using any kind of clustering, the measurement of proximity becomes a concern. The measurement of proximity can be either a geometric distance or a similarity relation defined between terms/concepts. Hence our proximity is of "similarity type" where the larger the similarity value of two observations e.g., x_i and x_j (letting the data repository to be defined as $X=\{x_1, x_2, ...,x_n\}$), the closer they are. If the similarity denoted by $Sim(x_i, x_j)$ is equal to 1, then x_i and x_j are same, what in context of distance-based measurements can be interpreted as $Dis(x_i, x_j)=0$. Distances between synsets are derived from their similarities using the formula, *distance = (1.0 – similarity)*. We use dissimilarity and distance conveying the same meaning.

In this work we use Hierarchical Agglomerative Clustering (HAC) for discovering the number of dimensions (or senses) from a group of synsets. So we need to analyze proximity measures for clusters as well as synsets.

If C_1 and C_2 are two different clusters, we would indicate the similarity between the two clusters as $Sim(C_1, C_2)$. We refer to this interclass similarity. Assume that at some point of clustering process we have q clusters, where C_k denotes set of k clusters, with $k=1$ to q. If merging is essential, clusters $C_{i'}$ and $C_{j'}$ can be selected such that $i' \neq j'$ and $Sim(C_{i'},C_{j'})=Max_{ij}(Sim(C_i,C_j))$, where C_i are the clusters/concepts inside $C_{i'}$ and C_j are clusters inside $C_{j'}$. Two methods, which have been often used for calculating the distance between two clusters are the nearest and furthest neighbor rules [7]. The nearest neighbor rule defines the inter-cluster similarity as the similarity between the elements in each of the two clusters that are biggest:

$$Sim(C_1, C_2) = \underset{c_{1_i} \in C_1 \text{ and } c_{2_j} \in C_2}{MAX} (Sim(c_{1_i}, c_{2_j})) \qquad (1)$$

The furthest neighbor rule defines the inter-cluster similarity as the similarity between the elements of the two clusters that are smallest:

$$Sim(C_1, C_2) = \underset{c_{1_i} \in C_1 \text{ and } c_{2_j} \in C_2}{MIN} (Sim(c_{1_i}, c_{2_j})) \qquad (2)$$

For the measurement of proximity we can depend on information theoretic models of similarity [8]. Conceptual similarity between two concepts of a hierarchy can be judged by shared information of those concepts. The similarity between the concepts depends on the degree of shared information. The mutual information shared between two words X and Y are given by [9]:

$$I(X,Y) = log \frac{p(X,Y)}{p(X)p(Y)} \qquad (3)$$

where $p(X, Y)$ is the probability of seeing X and Y together in a corpus. However, if X and Y are both very common, then it is likely that they appear together frequently simply by chance and not as a result of any relationship between them. To reflect this, probability $p(X, Y)$ is divided by $p(X)p(Y)$, which is the probability that X and Y would appear together by

chance, if they were independent. Taking the logarithm of this ratio gives mutual information some desirable properties. For example, its value is respectively positive, zero, or negative according to whether *X* and *Y* appear together more frequently, as frequently, or less frequently than one would expect if they were independent.

According to Resnik [9], information content of a concept *C* is defined in the standard way, $log\frac{1}{p(C)}$ where *p(C)* is the probability of encountering an instance of *C* in a certain corpus. The probability is based on using the data corpus to perform a frequency count of all the words in the synset of concept *C* and in any synset of a descendent concept [8]. Now, the information shared by two concepts C_1 and C_2 is approximated by the information content of the lowest common ancestor C_3 that subsumes them in the hierarchy. Hence, similarity between C_1 and C_2 is given by:

$$Sim(C_1, C_2) = IC(C_3) = log\frac{1}{p(C_3)} \qquad (4)$$

where $IC(C_3)$ indicates the information content of C_3. Indeed, the probability is highly based on the corpus. There are other works on similarity measurement like [10]–[13]. Seco et al. [13] present a novel mechanism of measuring information content arguing that WordNet itself can be used to measure the metric for information content without the necessity of external resources (e.g., corpuses). Information content of a WordNet concept *C* is given as a function of the hyponyms it has. The information content derived from WordNet for *C* is as follows:

$$IC_{wn}(C) = \frac{log\left(\frac{hypo(c)+1}{max_{wn}}\right)}{log\left(\frac{1}{max_{wn}}\right)} = 1 - \frac{log\left(hypo\left(C\right)+1\right)}{log\left(max_{wn}\right)} \qquad (5)$$

where the function *hypo* returns the number of hyponyms of a given concept and max_{wn} is a constant that is set to the maximum number of concepts in the WordNet taxonomy. The denominator, which is equivalent to the value of the most informative concept, serves as a normalizing factor in that it assures that *IC* values are in the range [0, 1]. The formulation above guarantees that the information content decreases monotonically. Moreover, the information content of the top node of WordNet would yield an information content value of 0. WordNet 2.1 contains a total of 81426 noun concepts (i.e., noun synsets). Hence, in our work max_{wn}=81426.

Resnik [14] has formulated measurement of similarity using information content. We get similarity between two concepts C_1 and C_2 by the following formula:

$$Sim_{res}(C_1, C_2) = \underset{c \in S(C_1, C_2)}{Max} IC(C) \qquad (6)$$

where $S(C_1, C_2)$ is the set of concepts that subsumes C_1 and C_2. Seco et al. [13] derive another measure by applying linear transformation to Jiang and Conrath formula [24], transforming it into a similarity function. This measurement depends only on the IC values and we have already discussed a metric for information content that uses only WordNet statistics. The resulting formulation is:

$$Sim(C_1, C_2) = 1 - \left(\frac{IC_{wn}(C_1) + IC_{wn}(C_2) - 2 \times Sim_{res}(C_1, C_2)}{2}\right) \qquad (7)$$

To get the similarity between two words we have used this formula extracting the concepts to which they belong because equation 7 is suitable for concepts/synsets of WordNet. If S_1 is the set of synsets of a word w_1 and S_2 is the set of synsets of another word w_2 then we can calculate the similarity between w_1 between w_2 with the following formula:

$$Sim(w_1, w_2) = \underset{C_i \in S_1 \; And \; C_j \in S_2}{Max} (Sim(C_i, C_j)) \qquad (8)$$

Besides, we have used Wu-Palmer [23] semantic similarity method to find similarity between two synsets. Semantic similarity between pair of concepts C_1 and C_2 can be calculated by the Wu-Palmer semantic similarity method.

$$Sim(C_1, C_2) = 2 \times len(r, C_3) / (len(C_1, C_3) + len(C_2, C_3) + 2 \times len(r, C_3)) \qquad (9)$$

where C_3 is the lowest common ancestor between C_1 and C_2. The root is represented by the *r* and $len(r,C_1)$ represents the shortest path between the root (*r*) and concept C_1. The length between root (*r*) and C_3 is global between two concepts. The formula uses the depth of the common concepts for measuring the similarity between two synsets. Thus the increase in distance between *r* and C_3 decreases the distance between C_1 and C_2. For measuring similarity between two words using this approach we used the following formula:

$$Sim(w_1, w_2) = Sim(MaxCommonPath(S_1, S_2)) \qquad (10)$$

where *MaxCommonPath* is a function that returns two synsets C_1 and C_2 possessing maximum common path in the taxonomy such that $C_1 \in S_1$ and $C_2 \in S_2$. Now, $Sim(C_1, C_2)$ corresponds to equation 9.

B. Relative Distance Plane (RDP)

In this paper, we construct a multidimensional space to represent the proximity using relative distance plane (RDP) [18] for better visualization. We want to map related synsets near to each other. The formal algorithm for the plot is as follows:

1. Select any two synsets R_1 $(=S_i)$ and R_2 $(=S_j)$ from the set of synsets *S* as two reference points. Consider distance $dis(S_1, S_2)=d_{12}$.
2. For each synset S_m, $(m \neq i,j)$ of *S*, let us consider its distances from the two reference synsets as: d_{1m} and d_{2m}. The Euclidean *(X,Y)* coordinate in the RDP for all synsets S_m, $m=1, 2, ...,n-2, m \neq i, j$ can be generated as follows:

$$X[S_m] = \frac{(d_{12})^2 + (d_{1m})^2 - (d_{2m})^2}{2d_{12}}$$

$$Y[S_m] = \sqrt{(d_{1m})^2 - (X[S_m])^2}$$

For dimension retrieval, we used hierarchical agglomerative clustering [25] and average silhouette coefficient [22]. We only use the distance matrix for dimension retrieval. The use of RDP is not mandatory for this work, rather it is used only for better visualization.

C. Silhouette Coefficient

Assume that the cluster to which object i is assigned is denoted as A. Let $a(i)$ be the average dissimilarity of i to all other objects of cluster A. For any cluster C different from A, let $d(i,C)$ be the average dissimilarity of i to all objects of C. After computing $d(i,C)$ for all clusters C, the smallest one among them denoted as $b(i) = \min_{C \neq A} [d(i,C)]$ is selected. The silhouette coefficient of object i, $s(i)$, is then obtained by combining $a(i)$ and $b(i)$ as follows [22]:

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, & \text{if } a(i) > b(i) \end{cases}$$

Silhouette coefficient of a cluster is measured by the average silhouette of all the elements in the cluster [19]. An overall measure of goodness of a clustering can be obtained by computing the average silhouette coefficient of all points [20]. We measured the natural number of senses in a group of synsets by looking at the number of clusters at which there is a peak in the plot in the evaluation measure when it is plotted against the number of clusters.

III. Dimension Retrieval Algorithm

We propose a dimension retrieval mechanism for text data in this paper. In our approach, we form a dynamic method which is document set-based. Our aim is to construct the dimensions depending on the terms/keywords in the set of documents. Retrieval of senses for the dimensions executes in several steps. The steps are as follows:

Step 1. Select all keywords from the documents.

Step 2. Find corresponding synsets.

Step 3. Find coordinate synsets of the synsets found in *step 2*. We denote this set of synsets by S.

Step 4. Construct dissimilarity matrix using elements of S. This is a $|S| \times |S|$ matrix.

Step 5. Retrieve senses using hierarchical agglomerative clustering with the maximum average silhouette coefficient.

For the selection of keywords, we depended on the header of each of the documents. We consider that if the number of retrieved senses is n, then the number of preliminary dimensions in the multidimensional space equals n. We pick up random reference synsets at each dimension; from which newly arrived synsets in the repository can be plotted. For example, assume that there are two dimensions and the reference synsets in the dimensions are S_1 and S_2. Now, let us consider the newly arrived synset S is to be plotted in the two dimensional space. Dissimilarity, $d_1(S_1, S)$ and $d_2(S_2, S)$ can be presented as distances of S from S_1 and S_2. Let arc_1 be the arc with S_1 center and d_1 radius, arc_2 be the arc with S_2 center and d_2 radius. The intersection of arc_1 and arc_2 indicates the synset S. This is illustrated in Fig. 1.

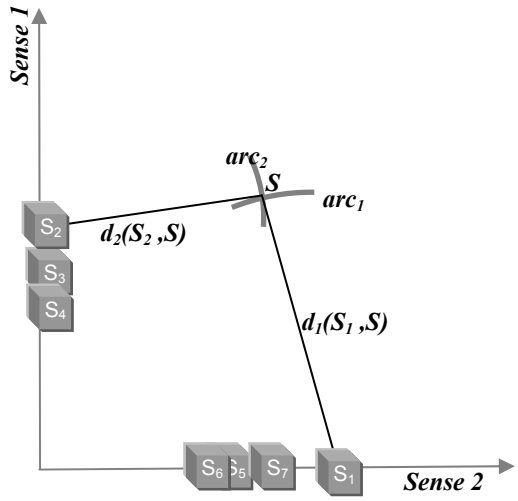

Fig. 1. Mapping S in multidimensional space.

Let us consider the set of unique keywords, $T=\{t_1, t_2, \ldots, t_n\}$ where t_i is a keyword and $1 \le i \le n$. We find all the synsets from WordNet where the keywords of T are included. Consider this set of synsets, $S'=\{S_1, S_2, \ldots, S_N\}$ where $N \ge n$ because a keyword can be included in more than one sysnset. We construct the WordNet-based but data specific hierarchy using S'. Elements of S' are the lowest level concepts of the tree. It should be mentioned that all upper level concepts are also synsets (i.e., WordNet concepts). Using this hierarchy, we can calculate similarities between all of the lowest level concepts. Hence we can calculate the metric for each synset of S'. It can be utilized to cluster synsets in such way that similar synsets are placed close together. The motivation

TABLE I
Comparison of Similarity Measures

	Word 1	Word 2	RG [16]	NS (Eq. 8)	WP (Eq. 10)
1	automobile	wizard	0.03	0.10	0.13
2	asylum	monk	0.10	0.08	0.17
3	glass	magician	0.11	0.20	0.29
4	boy	rooster	0.11	0.16	0.35
5	cushion	jewel	0.11	0.24	0.33
6	monk	oracle	0.23	0.22	0.53
7	boy	sage	0.24	0.22	0.62
8	automobile	cushion	0.24	0.31	0.40
9	furnace	implement	0.34	0.32	0.44
10	crane	rooster	0.35	0.50	0.67
11	crane	implement	0.59	0.39	0.60
12	oracle	sage	0.65	0.54	0.67
13	bird	crane	0.66	0.48	0.82
14	bird	cock	0.66	0.46	0.93
15	brother	monk	0.69	0.91	0.93
16	asylum	madhouse	0.76	0.85	0.93
17	cord	string	0.85	1.00	0.89
18	journey	voyage	0.90	0.77	0.91
19	autograph	signature	0.90	1.00	0.91
20	coast	shore	0.90	0.99	0.89
21	cushion	pillow	0.96	0.81	0.89
22	cemetery	graveyard	0.97	1.00	1.00
23	automobile	car	0.98	1.00	1.00
24	midday	noon	0.99	1.00	1.00
25	gem	jewel	0.99	1.00	1.00

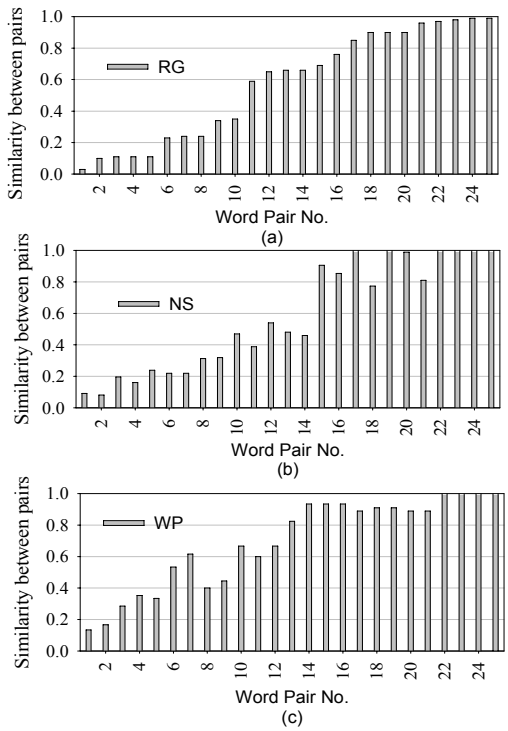

Fig. 2. Bar-charts for different kinds of similarity measurements of TABLE I. (Word Pair No. matches with the numbering used in TABLE I)

behind using the lowest level concepts is that the lowest level concepts have the highest information content value (*IC* of equation 5) and they compose the most informative elements of each document. As similarity is un-directional, we cannot use once single synset as reference to measure relatedness to the others. Hence to figure out the proximity between synsets we have to calculate similarity between all S_i and S_j where $1 \le i, j \le n$ and $i \ne j$ which have $O(n^2)$ complexity.

When evaluating our approach, we used two proximity measures (equation 7 and 9) for step 4 of the algorithm. To examine the correctness of our approaches, we compare them with human interpretations [16] in the next section. Besides, we provide a simple example illustrating the functionality of dimension retrieval algorithm in the appendix.

IV. Analysis on Measures of Proximity

TABLE I contains a comparison between our proximity measurements with human interpreted similarities. We selected 25 pairs of words from the list used in [16]. The column titled RG in TABLE I contains the normalized similarity values from [16]. All the similarity values in TABLE I have been normalized to [0, 1] range to make results comparable. NS and WP columns of the table contain outcomes of similarity measures between pairs of words. In TABLE I, we used equation 8 (NS) and 10 (WP) respectively. It should be noted that equation 8 and 10 provide similarity measurements between two words whereas equation 7 and 9 present similarity measures between two synsets. In the dimension retrieval algorithm proposed in section III, we constructed the dissimilarity matrix (*step 4*) utilizing equation 7 and 9.

TABLE II
Documents for the Experiment

Doc No.	Content
1.	***Silkworm***s are easy, fun and educational to grow in a classroom or at home. They are ***caterpillar***s that spin a ***silk cocoon*** and change into moths while inside. After hatching from an egg, the ***worm***s take one month to grow large enough to spin the ***silk***. They spend three weeks in the ***cocoon***, then emerge as a ***moth*** to mate and lay eggs. The eggs hatch into ***worm***s in a few weeks, and then the cycle continues.
2.	***Lepidoptera*** (*moth*s and butterflies) are the second most diverse pest *insect* order outnumbered only by the beetles. There is hardly any cultivated plant that is not attacked by at least one *lepidopteran* pest. As pollinators of many plants, adult moths and butterflies are usually beneficial insects that feed on nectar using their siphoning proboscis. The *caterpillar*s however almost always have chewing mouthparts that are suitable for feeding on various parts of a plant. Most *caterpillar*s are defoliators or miners of succulent plant tissues.
3.	***Jute*** is a long, soft, shiny ***plant fiber*** that can be spun into coarse, strong threads. It is produced from plants in the genus Corchorus, which see for botanical information and other uses. ***Jute*** is one of the cheapest ***natural fiber***s, and is second only to cotton in amount produced and variety of uses. It is much much cheaper than ***wool***. ***Jute fiber***s are composed primarily of the plant …
4.	….. i am about 45lbs below my ideal weight. i never follow organized sports. i don't believe in anything, not even nihilism. i have an aversion to public displays of affection. i believe the ***antichrist*** is a ***catholic***. i am practically blind without my glasses. i rock out to music from the 1930's. i am a star trek trivia ***nerd***. Sometimes people call call me ***insect***, because i am a generally ***unpleasant person*** because i am honest when its completely uncalled for and i have no life.
5.	***Jute*** is a member of a Germanic people who conquered England and merged with the ***Angle***s and ***Saxon***s to become ***Anglo-Saxons***. ….. Hengist is trained for the chieftainship. The Danes begin to encroach upon the lands of the ***Jute***s and ***Angle***s, and after intense debate, are allowed to settle. Hengist's first raid into Brittania. Hengist is married at a young age. King Wihtgil's death and deathbed injunctions to Hengist. Political strife at the Council of Elders. Hengist renounces the throne, which is given to Horsa. Hengist takes service with his friend and subordinate, King Hnaef.

Fig. 2 shows the corresponding bar-charts of TABLE I. It depicts that both our approaches (NS and WP) follow the trend of human interpretation (RG) although there are rare minor exceptions which in our opinion, are a result of WordNet-specific noun taxonomy. Hence the measure of proximity we use to construct the distance matrix for our dimension retrieval algorithm in section III, is proven to be close to human interpretation.

V. Experiments on Dimension Retrieval

In our experiment we used five documents and retrieved corresponding keywords from their headers. The documents are given in TABLE II. 20 keywords have been retrieved from these documents. The keywords are *silkworm*, *caterpillar*, *silk*, *cocoon*, *worm*, *moth*, *lepidoptera*, *insect*, *lepidopteran*, *jute*, *plant fiber*, *natural fiber*, *wool*, *antichrist*, *catholic*, *nerd*, *unpleasant person*, *angle*, *saxon* and *anglo-saxon*. These keywords are overlapped with a total of

36 synsets provided by WordNet. After incorporating coordinate synsets we found a total of 1378 synsets in the domain. We have applied our sense retrieval algorithm using two similarity measurement techniques which are portrayed in equation 7 and equation 9. Their corresponding RDP plots are illustrated in Fig. 3 (a) and (c) respectively, comparisons of average silhouette coefficients versus numbers of clusters are drawn in Fig. 3 (b) and (d). As the tendency of average silhouette coefficients is downward, the plots of (b) and (d) show average silhouette coefficients up to 50 clusters and the rests of the plots are omitted.

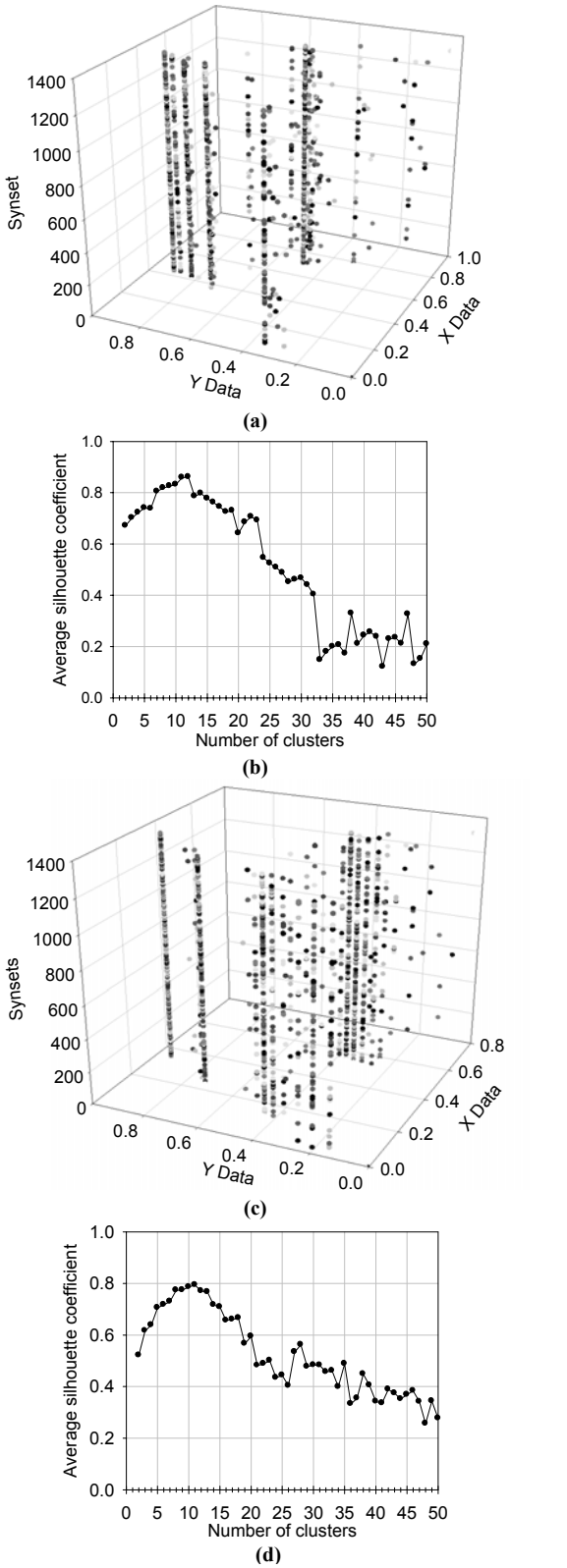

Fig. 3. Plot for RDP and average silhouette coefficient where similarities are measured by equation 7 (plot a and b) and equation 9 (plot c and d).

Fig. 3 (a) and (b) correspond to equation 7. The plot of Fig. 3 (b) shows that the maximum average silhouette coefficient is generated when there are 12 clusters. Fig. 3 (c) and (d) correspond to equation 9. Fig. 3 (d) shows that maximum average silhouette coefficient is produced when there are 11 clusters. Both of these similarity measurements produce close results. Experimental results show that there can be a total of 11 or 12 dimensions for the set containing 1378 synsets which is basically generated from 21 keywords.

VI. Conclusion

The major focus of our work presented in this paper is to reflect senses of keywords in dimensions with the aim of organizing descriptive data in multidimensional space with dimensionality lowered by aggregation of terms which have common meanings (i.e., are similar). The developed system measures proximity of terms derived from data corpus using an ontology, which represents background knowledge. It retrieves senses of related synsets occurring in the corpora, so that data can be organized in sense-based multidimensional space. Experimental results show that sense retrieval is possible using WordNet noun taxonomy as background knowledge. This work leads toward human-like organization of descriptive data. In future, we want to concentrate on further dimensionality reduction to reduce complexity and to overlap less density-dimension with comparatively higher density dimensions. All the experiments in this paper are done using WordNet noun taxonomy. We are interested in investigating few more ontologies in future to perform more thorough comparison with human-like organization of data.

Appendix

We provide here an example clarifying dimension retrieval algorithm. We consider a case where we have only a single

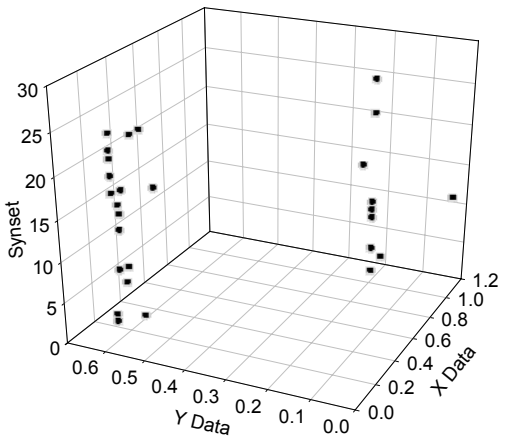

Fig. 4. All the synsets are placed in different planes and the using RDP plot for X and Y data.

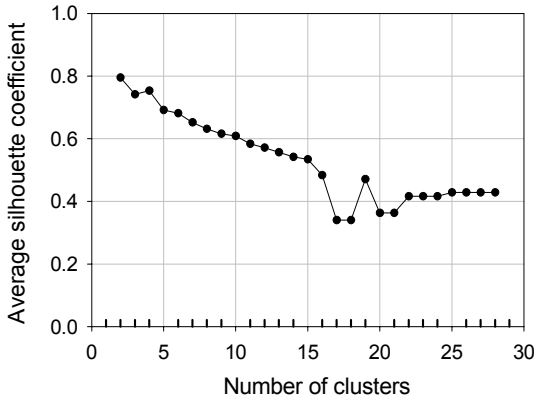

Fig. 5. Selection of natural number of senses by plotting average silhouette coefficients versus number of clusters

keyword *"lexicon"* in the corpora meta-data. *"lexicon"* is incorporated with two synsets in WordNet. 28 synsets were selected during the simulation including the coordinate synsets of the two corresponding synsets (step 3 of the algorithm). Then we construct 28×28 distance matrix using a similarity measure (equation 7 in this example). If we plot these synsets taking any two random synsets as reference points and measure their corresponding X and Y distances from the similarity matrix of synsets, we get the RDP plot (example drawn in Fig. 4).

Thereafter, we apply hierarchical agglomerative clustering to retrieve clusters from the synsets. But agglomerative clustering does not provide the natural number of clusters. Hence we calculate average silhouette coefficient for every number of clusters. The plot is given in Fig. 5.

Fig. 5. shows that maximum average silhouette coefficient is found when the number of clusters is two. This means, the natural number of clusters in the set of 28 synsets is 2. Now we look at all the synsets of TABLE III and TABLE IV. They contain two senses of *"lexicon"*. This shows that the technique reflects natural senses from a set of keywords, which can be used as dimensions of text repository.

TABLE III
SENSE 1 OF *LEXICON*

No.	Synset Description
1.	gazetteer
2.	unabridged dictionary
3.	desk dictionary, collegiate dictionary
4.	spell-checker, spelling checker
5.	Oxford English Dictionary, O.E.D., OED
6.	learner's dictionary, school dictionary
7.	pocket dictionary, little dictionary
8.	bilingual dictionary
9.	etymological dictionary

TABLE IV
SENSE 2 OF *LEXICON*

No.	Synset Description
1.	lexis
2.	place
3.	process, cognitive process, mental process, operation, cognitive operation
4.	inability
5.	public knowledge, general knowledge
6.	cognitive factor
7.	equivalent
8.	practice
9.	mind, head, brain, psyche, nous
10.	perception
11.	process, unconscious process
12.	vocabulary, lexicon, mental lexicon
13.	content, cognitive content, mental object
14.	episteme
15.	information
16.	history
17.	attitude, mental attitude
18.	ability, power
19.	structure

REFERENCES

[1] T. R. Gruber, "A translation approach to portable ontology specifications", Knowledge Acquisition, vol. 5, no..2, pp. 199–220, 1993.

[2] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. 1990. "Five papers on WordNet ", Technical report, Cognitive Science Laboratory, Princeton University, 1990.

[3] Nancy Ide and Jean Veronis. "Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art." Computational Linguistics, vol. 24, no. 1, pp. 1–40, 1998.

[4] Stevenson M., Wilks Y., **"**The interaction of knowledge sources in word sense disambiguation", Computational Linguistics, vol. 27, no. 3, pp. 321 – 349 (MIT press), 2001.

[5] A. Montoyo, M. Palomar, "Word Sense Disambiguation with Specification Marks in Unrestricted Texts", IEEE 11th International Workshop on Database and Expert Systems Applications (DEXA'00), pp. 103 –107, 2000.

[6] Chen MS, Han J, and Yu PS, "Data Mining: An Overview from a Database Perspective." , IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883, 1996.

[7] Yager, R.R., "Intelligent control of the hierarchical agglomerative clustering process", Systems, Man and Cybernetics, Part B, IEEE Transactions on , vol. 30, no.6 pp. 835- 845, Dec 2000.

[8] Cross V. Youbo Wang, "Semantic Relatedness Measures in Ontologies Using Information Content and Fuzzy Set Theory", The 14th IEEE International Conference on Fuzzy Systems, 2005. (FUZZ '05). pp. 114- 119, May 2005.

[9] Resnik P. "Selection and Information : A Class based Approach to Lexical Relationships", PhD. dissertation at the University of Pennsylvania. Technical Report IRCS-93-42, November 1993.

[10] Jay J. Jiang, David W. Conrath., "Semantic similarity based on corpus statistics and lexical taxonomy", Proceedings of International Conference on Research in Computational Linguistics, Taiwan, 1997.

[11] Sussna M. "Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network", Proceedings of the second international conference on Information and knowledge management, pp: 67 – 74, Washington, D.C. 1993.

[12] Dekang L., "An information-theoretic definition of similarity", In Proceedings of the 15th International Conference on Machine Learning, Madison, WI, 1998.

[13] Seco N., Veale T., and Hayes J., "An Intrinsic Information Content Metric for Semantic Similarity in WordNet", 16th European Conference on Artificial Intelligence 2004 (ECAI '04), pp. 1089-1090, 2004.

[14] Resnik P., "Using information content to evaluate semantic similarity in a taxonomy", Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448–453, 1995.

[15] Resnik P., "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language", Journal of Artificial Intelligence Research, vol. 11, pp. 95–130, 1999.

[16] Rubenstein, H. and J. B. Goodenough., "Contextual Correlates of Synonymy", Communications of the ACM, vol. 8, no. 10, pp. 627–633, 1965.

[17] Miller G.A. and Charles W.G., "Contextual Correlates of Semantic Similarity", Language and Cognitive Processes, vol. 6, no. 1, pp. 1-28, 1991.

[18] Somorjai R. L, Dolenko B, Demko A, Mandelzweig M, Nikulin AE, Baumgartner R, Pizzi NJ., "Mapping high-dimensional data onto a relative distance plane--an exact method for visualizing and characterizing high-dimensional patterns", Journal of Biomedical Informatics, vol. 37, no. 5, pp. 366-79, 2004.

[19] Adam N. R., Janeja V. P., Atluri V., "Neighborhood Based Detection of Anomalies in High Dimensional Spatio-temporal Sensor Datasets", ACM Symposium on Applied Computing, March 2004.

[20] Tan P. N., Steinbachm M., Kumar V., "Introduction to data mining", Published by Addison-Wesley, ISBN: 0321321367, pp. 539-547, April 2005.

[21] Banerjee, A. Dave, R. N., "Validating clusters using the Hopkins statistic", Proceedings of IEEE International Conference on Fuzzy Systems 2004, vol. 1, no. 1, pp. 149- 153, July 2004.

[22] Fu-ren Lin, Chih-ming Hsueh, "Knowledge map creation and maintenance for virtual communities of practice", Inf. Process. Manage. vol. 42, no. 2, pp. 551-568, 2006.

[23] Wu Z. , Palmer M., "Verb semantic and lexical selection", Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics, pp. 133–138, 1994.

[24] Jiang J. and Conrath D., "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", 10th International Conference on Research on Computational Linguistics (ROCLING X), 1997.

[25] Han J. and Kamber M. "Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers, pp. 408–411, 2006.

[26] Han J. and Kamber M. "Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers, pp. 72–86, 2006.

[27] Dhillon I. S., Guan Y., and Kogan J., "Iterative Clustering of High Dimensional Text Data Augmented by Local Search", Proceedings of the Second IEEE International Conference on Data Mining, Japan, pp. 131-138, December 2002.