**Supervised and Unsupervised Aspect CategoryDetection for Sentiment Analysis WithCo-Occurrence Data**

**Abstract**

Using online consumer reviews as electronic word of mouth to assist purchase-decision making has become increasingly popular. The Web provides an extensive source of consumer reviews, but one can hardly read all reviews to obtain a fair evaluation of a product or service. A text processing framework that can summarize reviewwould therefore be desirable. A subtask to be performed by such a framework would be to find the general aspect categories addressed in review sentences, for which this paper presents two methods. In contrast to most existing approaches, the first method presented is an unsupervised method that applies association rule mining on co-occurrence frequency data obtained from a corpus to find these aspect categories. While not on par with state-of-the-art supervised methods, the proposed unsupervised method performs better than several simple baselines, a similar but supervised method, and a supervised baseline, with an F1-score of 67%. The second method is a supervised variant that outperforms existing methods with an F1-score of 84%.

**Architecture**

****

**Existing System**

The lexicon directly associates aspects with categories. Sentences with no assigned category went through the postprocessing step, where the sentence was labeled with the category with maximum posterior probability. The lexicon learned from YELP data significantly improved the F1-score, which was reported to be 88.6% and ranked first among 21 submissions in SemEval-2014 workshop. However, for fair comparison, the score obtained without using the lexical resources derived from the YELP data, which is an F1-score of 82.2%, is reported in the evaluation, as our proposed supervised method to which it is compared also does not use external knowledge.

**Proposed System**

In this paper, both an unsupervised and a supervised method are proposed that are able to find aspect categories based on co-occurrence frequencies. The unsupervised method uses spreading activation on a graph built from word co-occurrence frequencies in order to detect aspect categories. In addition, no assumption has to be made that the implicit aspects are always referred to explicitly, like it is done. The proposed unsupervised method uses more than just the literal category label by creating a set of explicit lexical representations for each category. The only required information is the set of aspect categories that is used in the data set. The supervised method on the other hand uses the co-occurrences between words, as well as grammatical relation triples, and the annotated aspect categories to calculate conditional probabilities from which detection rules are mined.

**Future Work**

In terms of future work, we would like to investigate how injecting external knowledge would improve the results. While lexicons are a good way of doing that, we are especially interested in exploiting more semantic alternatives, like ontologies or other semantic networks. Also, As We Are Dealing With Unbalanced Data, We Plan To Explore Machine Learning Techniques that address this problem.

**Module Implementation**

1. **Removing Notational Words**

With the seed words known, the general idea of implicit aspect detection can be exploited to detect categories as well. The idea is to mine association rules of the form [notional word → category] from a co-occurrence matrix. Each entry in this co-occurrence matrix represents the frequency degree of two notional words co-occurring in the same sentence. Stop words, like the and, as well as less frequent words are omitted because they add little value for determining the categories in review sentences.

1. **Generating Association Rule**

Association rules are mined when a strong relation betweena notional word and one of the aspect categories exists, withthe strength of the relation being modeled using the co-occurrencefrequency between category and notional word. We distinguish between two different relation types: 1) *direct*and 2) *indirect* relations. A direct relation between two words*A* and *B* is modeled as the positive conditional probability*P(B*|*A)* that word *B* is present in a sentence given the fact thatword *A* is present. An indirect relation between two words *A*and *B* exists when both *A* and *B* have a direct relation witha third word *C*. This indicates that *A* and *B* could be substitutesfor each other, even though their semantics might notbe the same. Without checking for indirect relations, substitutesare usually not found since they do not co-occur oftentogether.

1. **Sentiment Analysis**

Sentiment analysis is a type of data mining that measures the inclination of people’s opinions through natural language processing (NLP), computational linguistics and text analysis, which are used to extract and analyze subjective information from the Web - mostly social media and similar sources. The analyzed data quantifies the general public's sentiments or reactions toward certain products, people or ideas and reveal the contextual polarity of the information.

1. **ClusteringResults**

As the name suggests, these models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. Also, the choice of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants.

**Algorithm**

1. **Spreading Activation Algorithm:** The data network structure used for the spreading activation algorithm will consist of vertices that represent the notional words, and links between two vertices representing a strictly positive co-occurrence frequency. Each link represents the direct relation between two notional words and receives weight equal to the conditional probability that word A co-occurs with word B, given that B appears in a sentence. This also means that the links receive direction as the conditional probability is not symmetric, making the data network structure a co-occurrence digraph.
2. **Clustering Algorithm:**Clustering analysis has been an emerging research issue in data mining due its variety of applications. With the advent of many data clustering algorithms in the recent few years and its extensive use in wide variety of applications, including image processing, computational biology, mobile communication, medicine and economics, has lead to the popularity of this algorithms. Main problem with the data clustering algorithms is that it cannot be standardized. Algorithm developed may give best result with one type of data set but may fail or give poor result with data set of other types. Although there has been many attempts for standardizing the algorithms which can perform well in all case of scenarios but till now no major accomplishment has been achieved. Many clustering algorithms have been proposed so far. However, each algorithm has its own merits and demerits and cannot work for all real situations.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV or Later Version

➢ RAM - 4 GB (min)

➢ Hard Disk - 40 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP or Later Version
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL

**Conclusion**

In this paper we have presented two methods for detectingaspect categories, that is useful for online review summarization. The first, unsupervised, method, uses spreadingactivation over a graph built from word co-occurrence data,enabling the use of both direct and indirect relations betweenwords. This results in every word having an activation valuefor each category that represents how likely it is to imply thatcategory. While other approaches need labeled training data to operate this method works unsupervised. The major drawbackof this method is that a few parameters need to be set beforehand,and especially the category firing thresholdsneed to be carefully set to gain a good performance. We havegiven heuristics on how these parameters can be set.The second, supervised, method uses a rather straightforwardco-occurrence method where the co-occurrencefrequency between annotated aspect categories and both lemmasand dependencies is used to calculate conditional probabilities.If the maximum conditional probability is higherthan the associated, trained, threshold, the category is assignedto that sentence.