E-Commerce Product Categorization

Srinivasu Gottipati and Mumtaz Vauhkonen

Introduction: Product classification for E-commerce sites is a backbone for successful marketing and sale of products listed on several online stores like Amazon, eBay, and craigslist etc. Since a large number of business users list their products and expect to find buyers for their products, it is crucial that the products are listed in accurate categories. This paper explores the experimental results that we conducted in using various forms of feature classification methods in combination with three main classifiers Naïve Bayes, SVM, K-Nearest Neighbors, along with LDA an unsupervised document topic classifier.

High level Steps followed for this classification process:

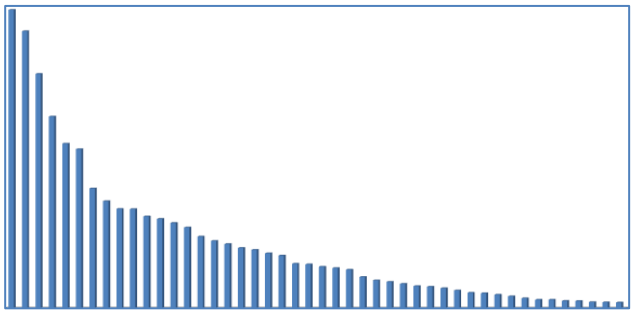
1. Data collection
2. Pre-processing
    a. Removal of less useful words like, of, the, an, in, and etc.
    b. Lower case conversion
3. Feature Selection and deriving unigram and bigram modals of the feature set using top occurring terms from each category, Info-gain, Chi square, and Latent Dirichlet Allocation (LDA).
4. Apply classification models. Naïve Bayes, Multi-Class SVM, K nearest neighbors (KNN) for both the unigram and bigram and combined unigram and bigram with a split of 50% with the data derived from each of the feature selection methods mentioned in step 3. These three models were selected with an intention to compare between generative(NB), discriminative(SVM) and non-parametric models(K-NN).
5. Analysis of the results

Section II: Relevant research discussion: With increasing speed of online marketing a sizable amount of research has been conducted. Several researchers have approached this problem from various angles and discussed their efficient outcomes. Young-Gon et al(1) implemented a modified Naïve Bayes model for product classification by applying a regular Naïve Bayes instead of text classifier and by making it treat each word as an attribute and making it accept weights assigned by the researchers. Though the accuracy is slightly high the main draw back in this approach is how to pick the right weight as it is based on observing the data and manually assigning the weights based on the features selected. Not choosing an appropriate weight would alter the results significantly. Lin and Shankar (2) have researched on using effective pre-processing techniques and multiclass features to increase accuracy in classification. Lee et al (3) discuss the classification process in terms of what exactly is classification in the context of multiple class relation models and they present a Semantic Classification model SCM. Wan and peng(4) used a fuzzy set modeling to identify the categories, but this model lacked the comparison of classification accuracies for evaluation. Meesad et al.(5) used chi-square as a classification method and compared to svm where SVM fared better but their research indicates potential for Chi-Square to be a robust classifier. The details are presented in the following section.

Section III: Implementation Model

Data Collection: To evaluate and test our approach to test which classifier with the given feature set would perform best in product classification, information on 35,000 products for 45 categories were gathered by crawling amazon site and scraping the pages to extract the attributes (title and description) . The category tree can be very deep and possibly contain many levels of depth. However, for the current scope of experiment, category tree was restricted to the top level and some of the things that could be placed into subcategories were placed on the top level. For example 'Luggage *Bags'* and 'Gym Bags' exist in the top level. The Attribute list consists of: *Title and Description.*

Category Selection: The graph in picture illustrates the distribution of the category classes. This reflects the typical distribution of long tail products e-commerce sites carry on the categories.



Below are top 5 categories with highest products:

Automotive	3000
Watches	2500
Shoes	2000
Electronics	1654
Bikes	1500

Table 1

Some other dense categories included Fitness, Musical instruments, Golf, Bikes, Sports Outdoor Accessories etc...

Data Pre Processing

For pre-processing, Apache Lucene libraries were used for tokenization, normalization, stop word removal, and stemming. As a first step in preprocessing, all the text was converted to lower case, and applied tokenization based on delimiters. (Tokenization can be very complex for multilingual scenarios and the current focus was restricted to English). Stop word removal was achieved

by manually crafting the list based on the various online resources. Porter-Stemmer algorithm implementation from Lucene was used to achieve stemming of the data.

Feature Selection

For all the products in the experiment, there were total of 45000+ unique unigrams and around 350,000+ unique bigrams. In order to reduce the number of features, feature selection was conducted by experimenting with various choices for bag of words by picking top words based frequency of occurrence from each category, Information Gain, Chi-Square Attribute evaluation and LDA top topic words. Unigrams and Bigrams were generated for top 100, 600, 1200, 4000, 10,000 feature sets and were tested as represented below in Table-2.

Chi-square Attribute Evaluation: The Chi-square attribute evaluation is based on the implementation of Pearson's χ^2 statistic. By applying this model for feature selection, the more unique a feature is the higher the χ^2 value . The formula is expressed as below and WEKA was used to run the chi-square attribute selection process.

$$\mathrm{X}^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

X^2 = Pearson's cumulative test statistic,

O_i = is observed frequency;

E_i = is expected (theoretical) frequency indicated by the Hypothesis

n = the number of cells in the table.

The following table illustrates the various combinations feature list were prepared (the cells that are checked are the ones against which the experiments were run):

Model	Frequency based	Info gain	Chi-square	LDA
Unigrams	✓	✓	✓	✓
Bigrams	✓			
Unigram + bigram (50-50 split)	✓	✓	✓	

Table - 2

Similarly for, BiGram modals, feature selection was conducted by picking top word pairs are that co-occurring in each of the categories and merging to get the final feature list. Once top X features are identified, then various experiments were carried out with varying number of features that included 100 features, 600 features, 1200 features, 4000 features and 10000 features. In addition to varying the number of features, separate experiments were done once by using only from title and other time using both title and description. The results of this analysis were detailed in Section IV of analysis section (We have seen consistently combination of title and description outperforming just tile alone, we have excluded analysis based on title in the analysis section).

LDA: Latent Dirichlet Allocation model was also used but primarily as a process to pick the top topic words and later using these topic words as unigrams feeding into the supervised classifiers to observe if it lead to increased accuracy..

Classification Model and Algorithms:

To identify the best way to increase accuracy in addition to feature selection, several classifiers were evaluated against the feature set and data collected. The classifiers that were used for the evaluation included multiclass SVM, Naïve Bayes (Multinomial), and K-nearest neighbors (with 5 neighbours)

Training and Test Sets

For training the classifier, 70% of the data set was used (24500 products out of 35,000) and for testing the rest of the 30% of the data set (10500 products) was used. From each product category of data 70% was used as a training data set and 30% as test data. The reason the percentages for each category were separately derived is to avoid the chance of over and under representation of data from a certain category.

Analysis of Results:

The results of the experiment with NB, SVM and K-NN give the following results:

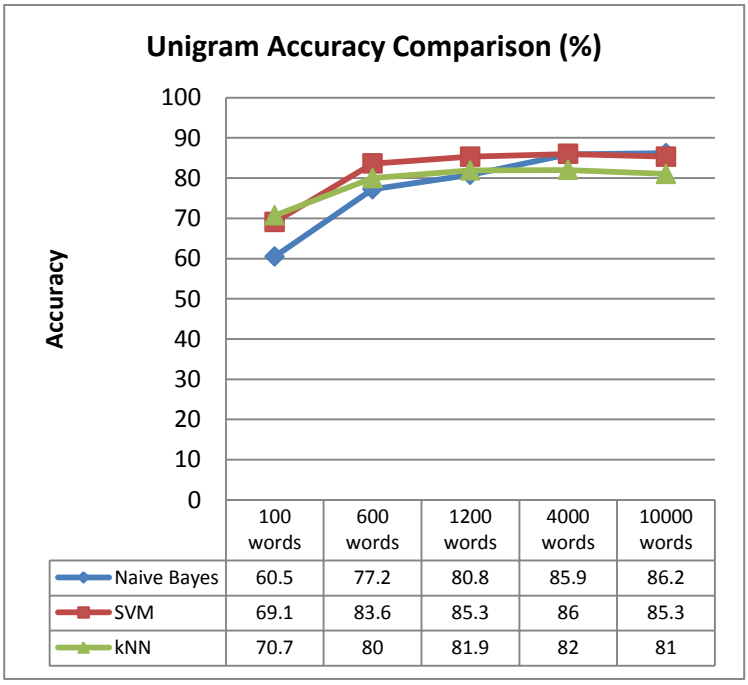
1. The unigrams as whole group outperform the bigrams for accuracy in classification.
2. Feature set size at 100 : When the feature set was small at 100 size, K-NN performed the best with all feature selection models. (refer to Figure 1 in next page) maintain accuracy of at least 1 to 2 percentage points above the rest. SVM trailed right behind K-NN with 1-2 percentage points below. However Naïve Bayes accuracy rates were much lower with a difference of 9 % points. A unique result set to is that, when the feature set of 100 derived using Chi-square model was used, it significantly increased the accuracy rates by more than 10 percentage points for Naïve Bayes and for SVM and K-NN at least by 4 percentage points. of all the classifiers as shown in Figure 2 on next page. The reason Chi-Square performed better was, it was able to identify words like "pedomet, trampoline rower sled" which have high degree of accuracy in associating with the respective category.
3. Performance at feature set size of 600 showed that all models got close to 2 or more % point boost using the chi-square model compared to the rest with SVM faring the best and KNN just trailing behind(refer to Figure -2 on next page). At the feature set size of 1200, Naïve Bayes started displaying higher accuracy gain compared to smaller data sets previously reaching 80% while SVM and KNN had steadily improved their accuracy rate till reaching 4000 feature set size with all models.

4. **Optimal Feature set size** seems to be at 4000 set size where the performance gap between regular unigram choice based on frequency and info-gain and Chi-square fared almost comparably with frequency based unigram model showing the best results for all classifiers.
5. **LDA** also showed decent levels of accuracy at 4000 data set but frequency-based and Chi-square models surpassed LDA. Refer to Table -3 below.
6. **Naïve Bayes** remained almost flat at the same levels as 4000 feature set size when the feature set size increased to 10,000 and surpassed all other models at 1000 features.
7. **SVM** steady improvement with increased feature set size till the point of 4000 and started a downward curve at 10,000 feature set size, the reason is that it started suffering from over fitting.
8. **KNN** was the worst performing as the feature set size increased.
9. **Running time** for `Naïve Bayes for fastest for all feature set sizes up to 10,000 with 1-2 minutes. SVM approximately max 15 minutes. K-NN ranged from 30min to 5 hours depending on the feature set size.

Three key points stand out from the experiments:

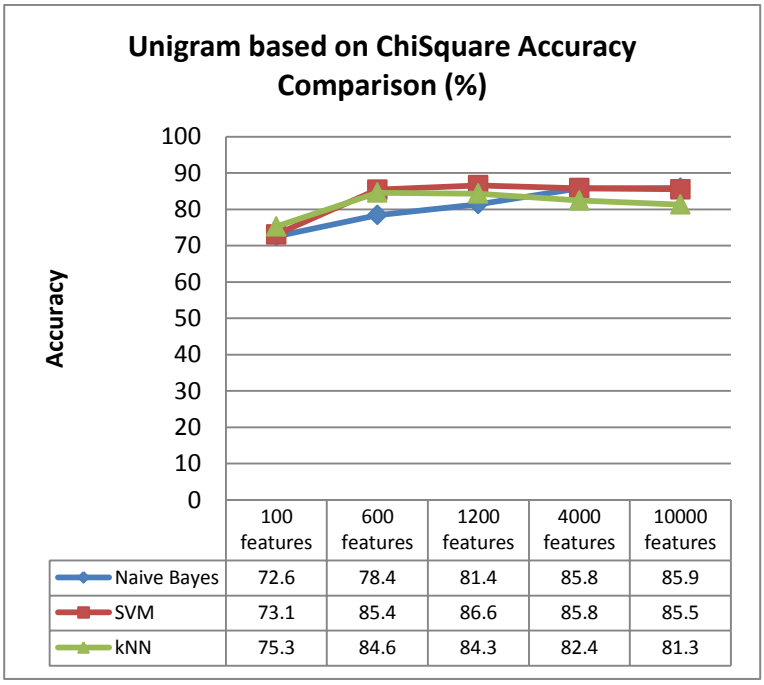
- Chi-square model showed a significant boost to all models in accuracy for a small feature set with Naïve Bayes getting the best boost. Though KNN performance was the best with a small feature set, the running time was significantly higher.
- For a large number of features, frequency based feature set of Unigrams gave best results for Naïve Bayes followed by SVM and KNN. Chi-square performance was also very similar to these results.
- Bi-gram models by themselves fared poorly compared to unigram model.

Presented below is the data and graphs that illustrate accuracy achieved by running the feature sets based on various ways from each feature extraction model (frequency based, infogain, Chi-square, LDA). All the graphs are listed below and next page to make it easy for comparing values. For kNN, 5 neighbors were used.



	100 words	600 words	1200 words	4000 words	10000 words
Naive Bayes	60.5	77.2	80.8	85.9	86.2
SVM	69.1	83.6	85.3	86	85.3
kNN	70.7	80	81.9	82	81

Figure 1 - Unigram Frequency Model



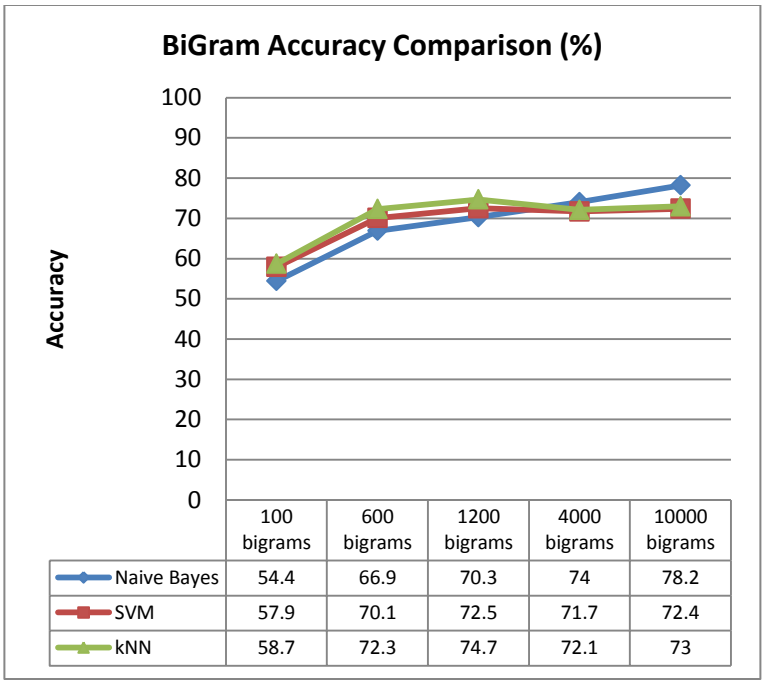
	100 features	600 features	1200 features	4000 features	10000 features
Naive Bayes	72.6	78.4	81.4	85.8	85.9
SVM	73.1	85.4	86.6	85.8	85.5
kNN	75.3	84.6	84.3	82.4	81.3

Figure 2 - Chi-square Unigram Model

LDA Topic Modeling: Topics modeled: 50, Features: 4000 (picked top 80 contributed from each topic)

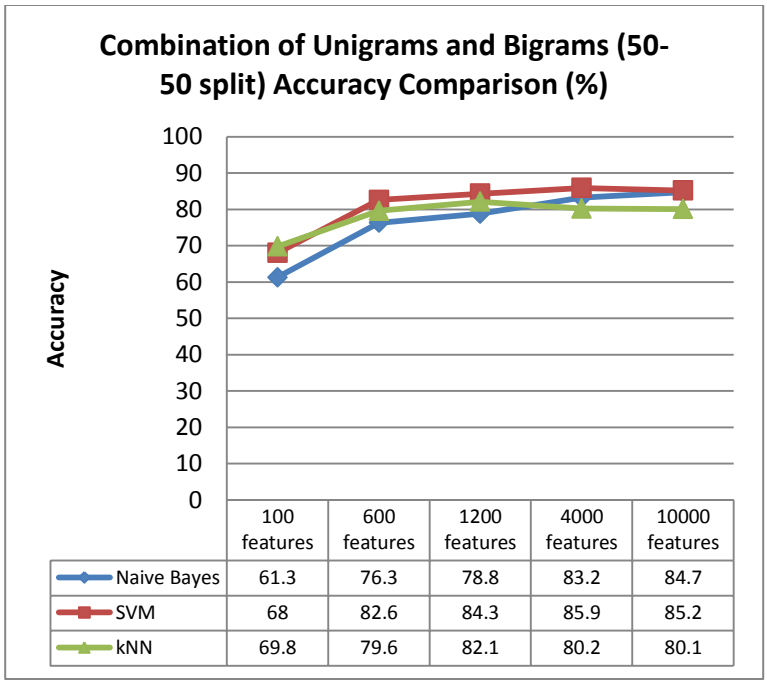
Naive Bayes	83.1
SVM	84
kNN	80.9

Table - 3



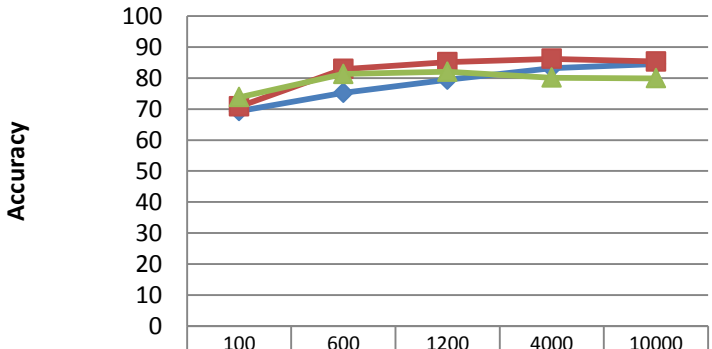
	100 bigrams	600 bigrams	1200 bigrams	4000 bigrams	10000 bigrams
Naive Bayes	54.4	66.9	70.3	74	78.2
SVM	57.9	70.1	72.5	71.7	72.4
kNN	58.7	72.3	74.7	72.1	73

Figure 3 -Bigram Frequency based Model



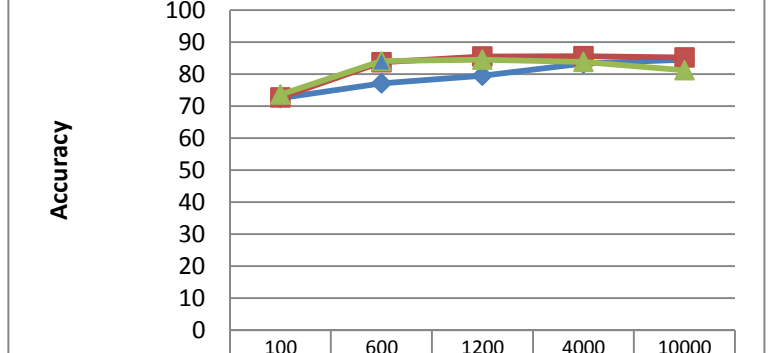
	100 features	600 features	1200 features	4000 features	10000 features
Naive Bayes	61.3	76.3	78.8	83.2	84.7
SVM	68	82.6	84.3	85.9	85.2
kNN	69.8	79.6	82.1	80.2	80.1

Figure 4 -Unigram and Bigram Split mode



	100 features	600 features	1200 features	4000 features	10000 features
Naive Bayes	69.3	75.2	79.4	83.2	84.5
SVM	70.8	82.9	85.1	86.2	85.3
kNN	73.8	81.3	82	80.1	79.9

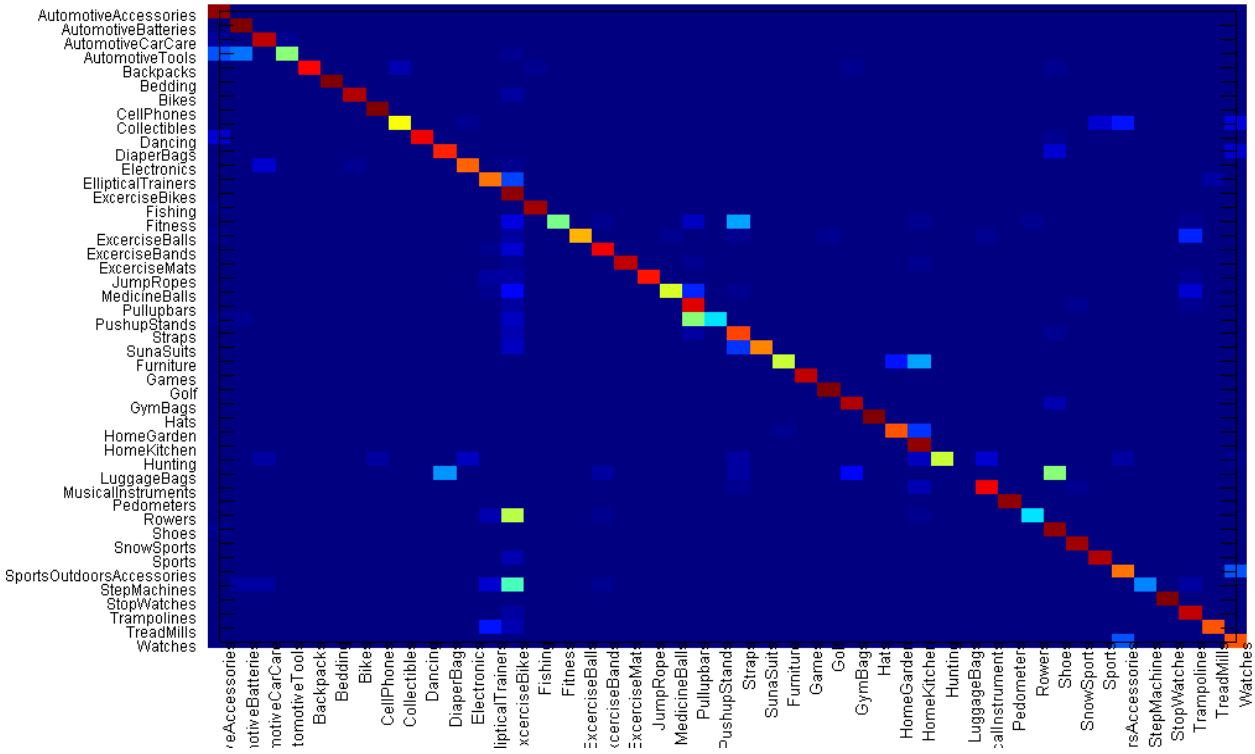
Figure 5 - Info Gain Unigrams and Bigrams Model



	100 features	600 features	1200 features	4000 features	10000 features
Naive Bayes	72.5	77.1	79.5	83.3	84.5
SVM	72.7	83.7	85.5	85.6	85.2
kNN	73.6	84.1	84.5	83.8	81.2

Figure 6 – Chi-Square Model Unigram-bigram split

Below is the confusion matrix derived from the best performing Chi-Square feature selection unigram modal with 4000 features based on SVM classifier:



As seen in the above confusion matrix, classifier was getting confused while classifying the items in 'Sports Outdoor Accessories' with the items in' Watches' category. The reason is that, 'Sports Outdoor Accessories' have "sporty watch items" and as watches category is dominant in training, classifiers were trying to maximize towards watch category. We have found similar patterns with other categories as well, where the product classification could fall into more than one category. This prompts for further research into multiple class relations.

Conclusion: The results of the experiments clearly indicate that with a small feature size set a Chi-square model of feature selection gives a significant boost to the classifiers. However at large feature set size, Naïve Bayes seems to gain the most accuracy with plain frequency based Unigram model for feature extraction and LDA fared at an average level. Further research needs to be done in refining the above product classification approach that can include more inputs such as images, tech specification etc. with an added functionality of creating a new category based on the incremental learning. In addition, a formal approach needs to be explored to suggest multiple categories where the classifiers have confusion among various categories and further research needs to be conducted in extending this to deal with nested category hierarchy.

References

1. **Modified naïve bayes classifier for e-catalog classification.** Kim, Young-Gon (School of Computer Science and Engineering, Center for E-Business Research, Seoul National University, Seoul 151-742, Korea, Republic of); Lee, Taehee; Chun, Jonghoon; Lee, Sang-Goo **Source:** *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v 4055 LNCS, p 246-257, 2006, *Data Engineering Issues in E-Commerce and Services - Second International Workshop, DEECS 2006, Proceedings*
2. **Applying Machine Learning to Product Categorization Lin.I and Shankar S**. Stanford University. CS229.
3. **A semantic classification model for e-catalogs.** Kim, Dongkyu (Corelogix, Inc., SNU ICT, 138-516 Seoul, Korea, Republic of); Lee, Sang-Goo; Chun, Jonghoon; Lee, Juhnyoung Source: *Proceedings - IEEE International Conference on E-Commerce Technology, CEC 2004*, p 85-92, 2004, *Proceedings - IEEE International Conference on E-Commerce Technology, CEC 2004*
4. **A technique of e-commerce goods classification and evaluation based on fuzzy set**. Wan, Hongxin (Mathematics and Computer Science College, Jiangxi Science and Technology Normal University, Nanchang, China); Peng, Yun Source: *International Conference on Internet Technology and Applications, ITAP 2010 - Proceedings*, 2010, *International Conference on Internet Technology and Applications, ITAP 2010 - Proceedings*
5. **A chi-square test for word importance differentiation in text classification. Meesad, P; Boonrawd P; Nuipian V; 2011- 2011 International Conference on Information and Electronics Engineering IPCSIT Vol.6 2011.**