Decision Tree Approach for Consumer Health Information Search

D. Thenmozhi
Department of CSE
SSN College of Engineering
Kalavakkam, Chennai
theni_d@ssn.edu.in

P. Mirunalini
Department of CSE
SSN College of Engineering
Kalavakkam, Chennai
miruna@ssn.edu.in

Chandrabose Aravindan
Department of CSE
SSN College of Engineering
Kalavakkam, Chennai
aravindanc@ssn.edu.in

ABSTRACT

Health information search (HIS) is the process of seeking health related information on the Internet by public health professionals and consumers. Abundance of health related information on the Internet may help a consumer for self-management of illness. Present day search engines retrieve information on consumer queries, but all of the retrieved information may not be relevant to the given query. It is a challenging task to identify the relevant information for a query from the result. In this paper, we present our methodology for a task to identify whether the information available are relevant or irrelevant for a given query using a machine learning approach. The lexical features that are extracted from the text are used by a classifier to predict whether the text are relevant or not for the query. We have also included a statistical feature selection methodology to select the significantly contributing features for the classification. We have evaluated our two variations using the data set given by CHIS@FIRE2016 shared task. The performance is measured in terms of accuracy and we have obtained overall accuracy of 75.87% for the method without feature selection and 78.1% for the method using χ^2 feature selection. Statistical t-tests confirm that feature selection has significantly reduced the sizes of the models without affecting the performance.

Keywords

Consumer Health Information Search; Machine Learning; Classification; Decision Tree; Feature Selection

1. INTRODUCTION

Information retrieval (IR) is the process of obtaining information relevant to a given query from a collection of resources. Internet is the major source of retrieving information for all domains. Health care is one of the domains where public health professionals and consumers seek for information from the Internet. Consumer Health Information Search (CHIS) is the process of retrieving health related information from Internet by common people to make some health related decisions and for self-management of diseases. Survey on CHIS have been is reported by Cline et al. [2], Zhang et al. [22] and Fiksdal et al. [3]. They have analyzed diverse purposes and diverse users on CHIS. Goeuriot et al. [4] analyzed the CHIS users based on varying information needs, varying medical knowledge and varying language skills. The existing search engines retrieve information based on keywords resulting in a large number of irrelevant information which may not satisfy diverse users of CHIS. The retrieval performance may be improved either by assisting the consumers to reformulate the query with more precise and domain specific terms [20, 13, 18], or by categorizing the retrieved information into relevant or irrelevant [9]. In this work, we have focused on the shared task of CHIS@FIRE2016 [12] which aims to identify text as relevant or irrelevant for a query. CHIS@FIRE2016 is a shared Task on Consumer Health Information Search (CHIS) collocated with the Forum for Information Retrieval Evaluation (FIRE). The goal of CHIS track is to research and develop techniques to support users in complex multi-perspective health information queries[1]. This track has two tasks. Given a CHIS query, and a document associated with that query, the first task is to classify whether the sentences in the document are relevant to the CHIS query or not. The relevant sentences are those from that document, which are useful in providing an answer to the query. The second task is to further classify the relevant sentences as supporting the claim made in the query, or opposing the claim made in the query. Our focus is on the first task of CHIS@FIRE2016.

2. RELATED WORK

Several research have been carried out in consumer health information search (CHIS) in recent years. Researchers analyzed the behaviour of the CHIS users [2, 22, 3] and the issues in searching for information [4]. The query construction, query reformulation and ranking of search result may improve the performance of CHIS. This section reviews the related work for CHIS.

2.1 Query Reformulation

Many researchers have analyzed the behaviour of the user in CHIS which help to reformulate the query for improving the performance of the retrieval. Zeng et al. [19] analyzed the query terms based on the query length, presence of stop words and frequency distribution and characterized the query as short and simple. Hong et al. [5] analyzed HealthLink search logs to find the behaviour of the user and found that the average length of queries submitted was 2.1 words. They have suggested that using of retrieval feedback may improve the consumer health information search performance. Spink et al. [14] analyzed the query logs of Alltheweb.com and Excite.com commerical web search engines to find the behaviour of health care users. They have reported that the average length of queries was 2.2 words.

[1]https://sites.google.com/site/multiperspectivehealthqa/home

Several researchers analyzed how consumers try to reformulate queries to improve the search performance. Toms and Latter [17] reported that consumers follow trial-and-error process to formulation of queries. Sillence et al. [11] stated that the queries are reformulated using Boolean operators by the consumers to alter search terms.

Several researchers presented algorithms for reformulating queries to improve health information search. Zeng [20] recommended additional query terms by computing the semantic distance among concepts related to the user's initial query based on concept co-occurrences in the medical domain. Soldaini et al. [13] proposed a methodology to bridge the gap between layperson and expert vocabularies by providing appropriate medical expressions for their unfamiliar terms. The approach adds the expert expression to the queries submitted by the users which they call as query clarifications. They have used a supervised approach to select the most appropriate synonym mapping for each query to improve the performance. Keselman et al. [7] supported the users with query formulation support tools and suggesting additional or alternative query terms to make the query more specific. They also educate the consumers to learn medical terms by providing interactive tools. Yunzhi et al. [18] proposed a methodology for query expansion using hepatitis ontology. They compute semantic similarity using ontology for finding the similarity of retrieval terms to improve retrieval performance.

2.2 Machine Learning Approaches for Health Information Search

Several researchers used machine learning approaches in health information search. Zhang et al. [21] used a machine learning approach for rating the quality of depression treatment web pages using evidence-based health care guidelines. They have used Naïve Bayes classifier to rate the web pages. Nerkar and Gharde [9] proposed a supervised approach using support vector machine to classify the semantic relations between disease and treatment. The best treatment for Disease is identified by applying voting algorithm. Automatic mapping of concepts from text in clinical report to a reference terminology is an important task health information search systems. Casteno et al. [1] presented a machine learning approach to bio-medical terms normalization for which they have used hospital thesaurus database.

Many works have been reported on query construction and query reformulation to improve the performance of consumer health information search. However, very few works have been reported on categorizing the retrieved information into relevant or irrelevant. Our focus is to categorize the information into relevant or irrelevant for the given query using machine learning approach in health care domain.

3. PROPOSED APPROACH

We have implemented a supervised approach for this CHIS task. The steps used in our approach are given below.

- Preprocess the given text
- Extract features for training data
- Build a model using a classifier from the features of training data
- Predict class label for the instance as "relevant" or "irrelevant" using the model

The steps are explained in detail in the sequel.

3.1 Feature Extraction

The given text is preprocessed before extracting the features by removing punctuations like ", ", –, ', ', and _and by replacing the term such as n't with not, & with and, 'm with am, and 'll with will. The terms of the each sentence in the given training text are annotated with parts of speech information such as noun, verb, determiner, adjectives and adverbs. In general, keyterms/features are extracted from the noun information. However, in medical domain, adjectives may also be contributed to the keyterms. For example, the sentence "Skin cancer is more common in people with light colored skin who have spent a lot of time in the sunlight." is relevant to the query "skin cancer". In this sentence, the adjective "light colored" is also important along with the nouns namely cancer, skin and sunlight to identify the sentence as relevant. Hence, all the nouns and adjectives from training data are extracted as features. We have considered all forms of nouns (NN^*) namely NN, NNS and NNP, and all forms of adjectives (JJ^*) JJ, JJR and JJS to extract the features. The extracted terms are lemmatized to bring them to their root forms. The feature set is constructed by eliminating all duplicate terms from the extracted terms.

We have used machine learning approach with two variations to identify whether the given text is relevant or not. The variations are

1. Approach without feature selection
2. Approach using χ^2 feature selection

The two variations are described in the following sub sections.

3.2 Approach without Feature Selection

We have used machine learning approach by extracting the linguistic features without explicit feature selection to build a model.

The set of extracted features along with the class labels namely relevant and irrelevant from training data are used to build a model using a classifier. We have used a decision tree based classifier called J48 to build the model. J48 classifier uses C4.5 algorithm to represent classification rules [10]. With J48 a model is constructed as tree during the learning phase.

The features are extracted for each instance of test data with unknown class label "?", similar to training data using the features vector of training data. The class label either "relevant" or "irrelevant" is predicted for the test data instances using the built model.

3.3 Approach using χ^2 Feature Selection

The number of features extracted by the methodology may be more. All of them may not be helpful to classify the text as "relevant" or "irrelevant". We have used a methodology which computes chi-square value for selecting the features from linguistic features. This χ^2 method selects the features that have strong dependency on the categories by using the average or maximum χ^2 statistic value.

Since, we have only two categories, we form a 2x2 feature-category contingency table which is called as CHI table for

every feature f_i. This table is used to count the co-occurrence observed frequency (O) of f_i for every category C and $\neg C$. Each cell at position (i, j) contains the observed frequency O(i, j), where $i \in \{f_i, \neg f_i\}$ and $j \in \{C, \neg C\}$. Table 1 shows 2x2 feature-category contingency table in which, $O(f_i, C)$ denotes the number of instances that contain the feature f_i belong to category C, $O(f_i, \neg C)$ denotes the number of instances that contain the feature f_i and are in not in category C, $O(\neg f_i, C)$ denotes the number of instances that does not contain the feature f_i but belong to category C, and $O(\neg f_i, \neg C)$ denotes the number of instances that neither contain the feature f_i nor belong to category C.

Table 1: Feature-Category Contingency Table

	C	$\neg C$
f_i	$O(f_i, C)$	$O(f_i, \neg C)$
$\neg f_i$	$O(\neg f_i, C)$	$O(\neg f_i, \neg C)$

The expected frequencies (E) for every feature f_i when they are assumed to be independent can be calculated from the observed frequencies (O). The observed frequencies are compared with the expected frequencies to measure the dependency between the feature and the category. The expected frequency E(i, j) is calculated from the observed frequencies (O) using the equation

$$E(i,\ j) = \frac{\Sigma_{a \in \{f_i, \neg f_i\}} O(a, j) \Sigma_{b \in \{C, \neg C\}} O(b, j)}{n} \quad (1)$$

where i represents whether the feature f_i is present or not, j represents whether the instance belongs to C or not, and n is the total number of instances.

The expected frequencies namely $E(f_i, C)$, $E(f_i, \neg C)$, $E(\neg f_i, C)$ and $E(\neg f_i, \neg C)$ are calculated using the above equation. Then the χ^2 statistical value for each feature f_i is calculated using the equation

$$\chi^2_{stat} f_i = \Sigma_{i \in \{f_i, \neg f_i\}} \Sigma_{j \in \{C, \neg C\}} \frac{(O(i,j) - E(i,j))^2}{E(i,j)} \quad (2)$$

The set of features whose χ^2_{stat} value is greater than $\chi^2_{crit(\alpha=0.05, df=1)} : 3.841$ are considered to be significant features and those features are selected for building a model using a classifier. The process to select χ^2 features from the linguistic features is given in Algorithm 1.

The model M_{chi} for the classification is build from training data by considering the selected feature set F_{chi} instead of F. The class label either "relevant" or "irrelevant" is now predicted for the test data instances by considering the built model M_{chi}

4. IMPLEMENTATION

We have implemented our methodologies in Java for the Shared Task on Consumer Health Information Search (CHIS): Task 1. The data set used to evaluate the task consists of five queries and a set of training data and test data for each query. The queries, number of training instances and number of test instances are given in Table 2.

4.1 Approach without Feature Selection

We have annotated the given sentences using Stanford POS tagger[2] which uses Penn Treebank tag set. For example, for the sentence "Skin cancer is more common in people with light colored skin who have spent a lot of time in the sunlight.", Stanford POS tagger annotate the sentence as "Skin_NN cancer_NN is_VBZ more_RBR common_JJ in_IN people_NNS with_IN light_JJ colored_VBN skin_NN who_WP have_VBP spent_VBN a_DT lot_NN of_IN time_NN in_IN the_DT sunlight_NN". All forms of nouns and adjectives are considered as features. In this example, "skin, cancer, common, people, light, time, sunlight" are extracted as features. Then the features are lemmatized. We have used Stanford lemmatizer to bring the features to their root form. Likewise, the features are extracted from all the training instances. Duplicates are eliminated to obtain a set of features for building a model. The number of features extracted for each query by this method is given in Table 4.

We have used J48 as a classifier to build the model with the extracted features. To implement the classifier, we have used Weka API[3]. Since Weka reads the feature vectors in "arff" format, we have prepared the feature vector files in "arff" format. The model is built by training the classifier using the training data feature vectors.

The class labels either "1" for "relevant" or "0" for "irrelevant" are predicted using the model for the test instances.

Algorithm 1 χ^2 Feature Selection

Input: Training data T, Set of linguistic features F
Output: Set of χ^2 features F_{chi}

```
Let Chi feature set F_chi = ∅
for (each f_i ∈ F) do
    for (each category C ∈ [relevant, irrelevant]) do
        Construct 2x2 feature-category contingency table (CHI table) with the observed co-occurrence frequencies (O) of f_i and C using T and F
        Calculate the expected frequencies (E) using CHI table
        E(i, j) = (Σ_{a∈{f_i,¬f_i}} O(a,j) Σ_{b∈{C,¬C}} O(b,j)) / n
        Calculate χ² value of f_i for C
        χ²_stat f_i = Σ_{i∈{f_i,¬f_i}} Σ_{j∈{C,¬C}} ((O(i,j)−E(i,j))²) / E(i,j)
    end for
    if χ²_stat f_i >= χ²_crit(α=0.05,df=1) : 3.841 then
        Add f_i to F_chi
    end if
end for
Return feature set F_chi
```

Table 2: Data Set for CHIS task

Query	Training	Test
Skin Cancer	341	88
E-Cigarettes	413	64
Vitamin-C	278	74
HRT	246	72
MMR-Vaccine	259	58

4.2 Approach using χ^2 Feature Selection

In this variation, we have selected set of features which significantly contribute to identify the classes, from the linguistic features. To select the features, we have used a sta-

[2]http://nlp.stanford.edu/software/tagger.shtml
[3]http://www.java2s.com/Code/Jar/w/Downloadwekajar.htm

tistical approach called χ^2 method. We have constructed the CHIS table for each feature f_i. For example, the CHIS table which shows the observed frequencies for the feature "estrogen", with respect to the query "HRT" is given in Table 3.

Table 3: CHIS Table for the feature "Estrogen" with respect to the query "HRT"

	Relevant	Irrelevant
Estrogen	39	14
¬Estrogen	167	26

The total number of training instances are 246 for the query "HRT". The expected frequencies are calculated from the CHIS table values using Equation 1. The expected frequencies obtained for the feature "Estrogen" are 44.0, 8.0, 161.0 and 31.0. The $\chi^2_{stat}(Estrogen)$ is computed using Equation 2 as 6.098236 which is greater than $\chi^2_{crit(\alpha=0.05,df=1)}$: 3.841. Thus, this "Estrogen" feature is selected as a candidate feature for building the model using the classifier. The number of features selected by this statistical method for all the queries given in the task are shown in Table 4.

Table 4: Number of features for the queries

Query	Without Feature Selection	χ^2 Feature Selection
Skin Cancer	742	31
E-Cigarettes	1014	36
Vitamin-C	715	25
HRT	547	10
MMR-Vaccine	751	12

Further, the feature vectors for the training data are constructed similar to our first approach in "arff" format and the model is built by J48 classifier using Weka API.

Table 5 shows size of the tree in terms of number of nodes which describe the model created for both variations of our approach. It is observed from Table 5 that the number of nodes used in the decision tree by J48 classifier is considerably reduced when χ^2 feature selection method is used.

Table 5: Size of the Tree

Query	Without Feature Selection	χ^2 Feature Selection
Skin Cancer	57	29
E-Cigarettes	51	27
Vitamin-C	21	3
HRT	23	7
MMR-Vaccine	39	3

To show that this reduction is statistically significant, we have applied a t-test on these 2 models. k-Fold paired t-test with one-tailed distribution is used to show that the reduction is significant when features are selected using χ^2. The $p-values$ obtained for size of the tree while applying paired t-test (one-tailed, 95% confidence) is 0.001236616 which is less than 0.05. This shows that the reduction in size of the tree is statistically significant.

The prediction is done for the test data as in our first approach to identify whether the test instances belong to one of the category "relevant" or "irrelevant".

4.3 Results

We have evaluated the performance of our methodologies using the metric accuracy. We have performed the 10-fold cross validation on training data. The cross validation accuracies given by the methodologies for the queries are summarized in Table 6.

Table 6: 10-fold cross validation accuracy (%) for the queries

Query	Without Feature Selection	χ^2 Feature Selection
Skin Cancer	92.96	85.34
E-Cigarettes	84.26	76.27
Vitamin-C	88.49	82.37
HRT	93.09	86.9
MMR-Vaccine	93.05	80.31

The performance of our both the methods on evaluating the test data is shown in Figure Table 7. It is observed from Table 7 that the accuracy obtained after χ^2 feature selection is more than the method without feature selection by 2.23%.

Table 7: Test data accuracy (%) for the queries

Query	Without Feature Selection	χ^2 Feature Selection
Skin Cancer	86.36	79.54
E-Cigarettes	65.25	64.06
Vitamin-C	73.0	78.38
HRT	87.5	87.5
MMR-Vaccine	67.24	81.03
Average Accuracy	75.87	78.1

We have compared our two approaches using k-fold paired t-test and Mcnemar test to show that the improvement in performance is statistically significant. We have applied 5-fold paired t-test (1-tailed, 95% confidence, 5 dataset) on our two approaches and we have obtained the $p-value$ of 0.278 for accuracy. Since, this $p-value$ is greater than 0.05, we can statistically infer that our approach using χ^2 feature selection does not reduce the performance of our system. When we apply Mcnemar test across all data sets, we obtain the $p-value$ of 0.5186 which is also greater than 0.05. These show that our feature selection approach significantly reduces the size of the model without compromising the performance.

5. CONCLUSIONS

We have presented a system for identifying whether the given text are relevant or irrelevant to a query. We have proposed two variations of our methodology namely an approach with all features and an approach with selected features based on chi-square statistical value. In both the methods, we have identified the features and feature vectors are constructed from training data. We have used J48 classifier to build a model with these feature vectors and the model is used to predict whether the test instances or "relevant" or "irrelevant" to the query. We have used the data set given by CHIS@FIRE2016 shared task to evaluate our methodology. We have performed a statistical t-test which shows our χ^2 feature selection method significantly reduces

the size of the model for CHIS@FIRE2016 data set. We have measured the performance of our approaches using the metric accuracy. We have obtained the accuracy of 75.87% and 78.1% for the method without feature selection and the method using χ^2 feature selection respectively for the Task 1 of CHIS@FIRE2016 shared task. Statistical t-tests namely k-fold paired t-test and Mcnemar test confirm that feature selection has significantly reduced the sizes of the models without affecting the performance. At present we have used parts of speech (POS) information and χ^2 value to extract and select the features respectively. Further, the features may be extracted based on the predicate information of the text [15, 16]. The CHIR value [8, 6] may be calculated from χ^2 value to select the features in future.

Acknowledgments

We would like to thank the management of SSN Institutions for funding the High Performance Computing (HPC) lab where this work is being carried out.

6. REFERENCES

[1] J. Castano, H. Berinsky, H. Park, D. Pérez, P. Avila, L. Gambarte, S. Benıtez, D. Luna, F. Campos, and S. Zanetti. A machine learning approach to clinical terms normalization. *ACL 2016*, page 1, 2016.

[2] R. J. Cline and K. M. Haynes. Consumer health information seeking on the internet: the state of the art. *Health education research*, 16(6):671–692, 2001.

[3] A. S. Fiksdal, A. Kumbamu, A. S. Jadhav, C. Cocos, L. A. Nelsen, J. Pathak, and J. B. McCormick. Evaluating the process of online health information searching: a qualitative approach to exploring consumer perspectives. *Journal of medical Internet research*, 16(10):e224, 2014.

[4] L. Goeuriot, G. J. Jones, L. Kelly, H. Müller, and J. Zobel. Medical information retrieval: Introduction to the special issue. *Inf. Retr.*, 19(1-2):1–5, April 2016.

[5] Y. Hong, N. de la Cruz, G. Barnas, E. Early, and R. Gillis. A query analysis of consumer health information retrieval. In *Proceedings of the AMIA Symposium*, page 1046. American Medical Informatics Association, 2002.

[6] M. Janaki Meena and K. Chandran. Naive bayes text classification with positive features selected by statistical method. In *In International Conference on Autonomic Computing and Communications, ICAC 2009*, pages 28–33. IEEE, 2009.

[7] A. Keselman, A. C. Browne, and D. R. Kaufman. Consumer health information seeking as hypothesis testing. *Journal of the American Medical Informatics Association*, 15(4):484–495, 2008.

[8] C. L. Li Yanjun and S. M. Chung. Text clustering with feature selection by using statistical data. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):641–652, 2008.

[9] B. E. Nerkar and S. S. Gharde. Best treatment identification for disease using machine learning approach in relation to short text. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 16(3):5–12, 2014.

[10] P. A. F. Pavel, Yonghong Peng and B. C. Soares. Decision tree-based data characterization for meta-learning. *IDDM-2002*, page 111, 2002.

[11] E. Sillence, P. Briggs, L. Fishwick, and P. Harris. Trust and mistrust of online health sites. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 663–670. ACM, 2004.

[12] M. Sinha, S. Mannarswamy, and S. Roy. CHIS@FIRE: overview of the CHIS track on consumer health information search. In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India, December 7-10, 2016*, CEUR Workshop Proceedings. CEUR-WS.org, 2016.

[13] L. Soldaini, A. Yates, E. Yom-Tov, O. Frieder, and N. Goharian. Enhancing web search in the medical domain via query clarification. *Inf. Retr. Journal*, 19(1-2):149–173, 2016.

[14] A. Spink, Y. Yang, J. Jansen, P. Nykanen, D. P. Lorence, S. Ozmutlu, and H. C. Ozmutlu. A study of medical and health queries to web search engines. *Health Information & Libraries Journal*, 21(1):44–51, 2004.

[15] D. Thenmozhi and C. Aravindan. An automatic and clause based approach to learn relations for ontologies. *The Computer Journal, Accepted for Publication, DOI: 10.1093/comjnl/bxv071*, 2015.

[16] D. Thenmozhi and C. Aravindan. Paraphrase identification by using clause based similarity features and machine translation metrics. *The Computer Journal, Accepted for Publication, DOI: 10.1093/comjnl/bxv083*, 2015.

[17] E. G. Toms and C. Latter. How consumers search for health information. *Health informatics journal*, 13(3):223–235, 2007.

[18] C. Yunzhi, L. Huijuan, L. Shapiro, and L. Travillian, Ravensara S.and Lanjuan. An approach to semantic query expansion system based on hepatitis ontology. *Journal of Biological Research-Thessaloniki*, 23(1):11, 2016.

[19] Q. Zeng, S. Kogan, N. Ash, R. Greenes, A. Boxwala, et al. Characteristics of consumer terminology for health information retrieval. *Methods of information in medicine*, 41(4):289–298, 2002.

[20] Q. T. e. a. Zeng. Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Association*, 13(1):80–90, 2006.

[21] Y. Zhang, H. Cui, J. Burkell, and R. E. Mercer. A machine learning approach for rating the quality of depression treatment web pages. *iConference 2014 Proceedings*, 2014.

[22] Y. Zhang, P. Wang, A. Heaton, and H. Winkler. Health information searching behavior in medlineplus and the impact of tasks. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 641–650. ACM, 2012.