electronic Journal of Health Informatics
http://www.ejhi.net
2009; Vol 4(1): e5

‘Qualities’ not ‘Quality’ – Text Analysis Methods to Classify Consumer Health Websites

Guocai Chen[1], Jim Warren[1,2], Joanne Evans[3]
[1] Department of Computer Science [2]School of Population Health, The University of Auckland, New Zealand

[3] Centre for Organisational and Social Informatics, Faculty of IT, Monash University, Australia

Abstract
There is an increasing need to help health consumers to achieve timely, differentiated access to quality online healthcare resources. This paper describes and evaluates methods for automated classification of consumer health Web content with respect to qualitative attributes relevant to the preferences of individual health consumers. This is illustrated in the context of identifying breast cancer consumer web pages that are ‘supportive’ versus ‘medical’ perspective, as compared to an existing manual classification employed by a breast cancer portal with personalised search preference options. Classification is performed based on analysis of word co-occurrences and an enhanced decision tree classifier (a decision forest). Current classification test results for ‘medical’ versus ‘supportive’ type resources are 90% accurate (95% confidence interval, 86-94%) using this decision forest classifier. These early results are indicating that language use patterns can be used to automate such classification with acceptable accuracy; however, a wider range of websites and metadata attributes needs to be assessed and compared to end-user feedback. Future application may be either in a tool to facilitate metadata coders in populating the databases of domain-specific portals such as BCKOnline, or in providing tagging or sorting on content type on live search results from health consumers.

Keywords: Consumer Health Information, Internet, Metadata, Natural Language Processing

1. Introduction

When confronted with a healthcare situation, people are increasingly turning to the Internet for information to aid in understanding diagnoses, deciding on treatment options and seeking psychosocial support for themselves, their family and their friends [1]. Vast quantities of health information are being made available online by a number of providers ranging from government agencies, pharmaceutical companies, commercial companies, charity organisations, community groups and individuals to service the information needs of medical professionals and healthcare consumers. As a result a keyword search using any of the major search engines on most healthcare topics will bring up thousands, hundreds of thousands, and even millions of hits of varying quality and relevance to a person’s particular health and life situation. The resulting information overload, where the amount of information exceeds a person’s ability to process it [2], can often add stress to an already stressful situation. Consequently there is much concern regarding how the quality, relevance, authority and accuracy of online information can be assessed in a timely manner by both healthcare consumers and medical professionals alike [3-4].

Many projects have been devised to address information overload and investigate ways in which timely, differentiated access to quality online healthcare resources can be provided. The provision of web portals, centred

The *electronic Journal of Health Informatics* is an international journal committed to scholarly excellence and dedicated to the advancement of Health Informatics and information technology in healthcare. ISSN: 1446-4381

on particular health topics and/or communities of users, is one such strategy [5-6]. The aim is to provide access to a reduced corpus of information resources that meet quality and relevance criteria. Portals can be further augmented by capturing and creating descriptive metadata about resources selected for inclusion. This structured, value-added information can then be used by portal users in searching, filtering, ranking, and in making judgements about what information is relevant to their needs and in which they wish to trust.

The Breast Cancer Knowledge Online Portal (BCKOnline) is an example of such an approach. Developed through collaboration between Monash University, BreastCare Victoria and the Breast Cancer Action Group, the portal provides a gateway to online information about breast cancer of relevance to breast cancer patients, their families, friends and carers. The portal incorporates metadata that describes relevant resources from a user-centred perspective [7]. Included in the description of resources is metadata about the type and style of information, the stage of breast cancer to which it relates, and the categories of users to which it applies [8]. The search interface allows portal users to indicate their information preferences along these lines. Usability studies show a high degree of satisfaction with BCKOnline [7], suggesting that it may be a useful model for information portals in the consumer healthcare sector.

The further development of the model underpinning BCKOnline as a generic approach to the provision of smart information portals is the subject of an Australian Research Council Discovery Project, "Smart Information Portals: Meeting the knowledge and decision support needs of health care consumers for quality online information." One of the key questions this project is addressing is how the metadata, that enriches the user experience and allows differentiated access to resources based on personal information needs, can be created in sustainable and scalable ways. Manual methods of metadata creation just cannot keep up with the vast quantities of healthcare information being made available online, and cannot easily respond to their increasing dynamism, complexity and volatility [9]. In addition those responsible for selecting resources for inclusion in a portal's knowledge repository of metadata descriptions need more sophisticated tools for discovering potential resources of relevance. In the case of BCKOnline, user information needs analysis identified the desire for more access to personal stories of breast cancer experiences, which are often buried deep in the result sets of the major search engines. Further development of the portal therefore requires investigation into how the generation of metadata describing relevant resources from a user-centred perspective can be automated.

2. Methods

A key distinction in BCKOnline is the resource type attribute, which identifies site content as any of 'medical' (evidence-based), 'supportive' (regarding support resources), and 'personal' (individual views). While it may be possible to contrive an ad hoc set of heuristics for distinguishing these classes of sites, we have focused on examining the language use as the basis for automated classification. Classically, a word frequency vector provides a set of features for classification, one feature for every distinct word used in any of the web pages under consideration. Depending on the details of the data pre-processing, such a method may yield some 5000 features for classification.

A more detailed treatment of language use is achieved using a method called Hyperspace Analogue to Language (HAL, [10]), where scores are accumulated based on the co-occurrence of words in proximity to one another, producing a much larger (but sparser) feature set. Burgess and Lund [11] demonstrated that HAL could be used to distinguish the emotional connotation of words. A HAL matrix is produced by passing a 'window' of a certain length over each word in a corpus and recording a score into the matrix for all words co-occurring within that window. Figure 1 illustrates the appearance of a HAL matrix produced with a window of size 3 passed over the text "Evaluating breast changes or masses usually starts with a mammogram or sonogram (ultrasound) performed by a radiologist." Words with minimal domain-specific semantics are removed in pre-processing. At the top of Figure 1 the window and its HAL score contribution is illustrated at the point where the window is applied to the word "breast."

2 0 2 1 0 …
Evaluating breast changes masses usually starts mammogram sonogram ultrasound performed radiologist

	evaluating	breast	changes	masses	usually	starts	mammogram	sonogram	ultrasound	performed	radiologist
evaluating	0	2	1	0	0	0	0	0	0	0	0
breast	2	0	2	1	0	0	0	0	0	0	0
changes	1	2	0	2	1	0	0	0	0	0	0
masses	0	1	2	0	2	1	0	0	0	0	0
usually	0	0	1	2	0	2	1	0	0	0	0
starts	0	0	0	1	2	0	2	1	0	0	0
mammogram	0	0	0	0	1	2	0	2	1	0	0
sonogram	0	0	0	0	0	1	2	0	2	1	0
ultrasound	0	0	0	0	0	0	1	2	0	2	1
performed	0	0	0	0	0	0	0	1	2	0	2
radiologist	0	0	0	0	0	0	0	0	1	2	0

Figure 1. Example HAL matrix for a short phrase.

We examine the classification of the 135 Supportive and 701 Medical type web pages indexed in the BCK-Online database using HAL features. In particular, we sample 80 each websites that are Supportive and 80 that are Medical. Using 8-fold cross-validation we assess classification accuracy on 10 test sites of each type held back from classifier training. We have previously reported classification results using a k-nearest neighbour classifier, AKLH ([12] with methods as per [13]) and with decision trees ([13]). In the case of decision trees, a tree is developed by a recursive process of selecting the word whose column in the HAL matrix best distinguishes the Supportive from the Medical cases, employing the ID3 algorithm (see http://en.wikipedia.org/wiki/ID3_algorithm; and [14]). As such, the trees are induced based on information gain (entropy reduction), selecting the best word (i.e., column) from the HAL matrix for each decision node in a recursive algorithm until all training data is correctly classified or the algorithm halts due to lack of a variable that can successfully discriminate. The resulting tree classifies new (test) cases by computing a HAL matrix on just the test Web page's corpus and then determining whether the cosine of HAL vectors of the words at decision nodes are larger when compared to Supportive or Medical training data at each node of the decision tree.

Herein we apply an enhanced decision tree based classifiers, using the concept of a decision forests (a set of independent decision trees that 'vote' on a solution; methods similar to [15]). Our decision forest is made up of a set of independent decision trees each based on a mutually-exclusive subset of the columns of the HAL matrix and induced by an ID3 algorithm as per above. The appropriate forest size (number of distinct trees) is unknown and so a range of sizes is explored.

Table 1. Highest-value components of HAL matrices for 80 Medical and 80 Supportive web sites.

medical	cancer	breast	women	treatment	patients	children	time	chemotherapy	risk	life
cancer	5200	15371	3570	2047	1387	40	249	925	2961	200
breast	15371	4448	3407	1516	1184	18	288	671	2496	212
women	3570	3407	2440	839	216	30	182	382	1220	223
treatment	2047	1516	839	1244	506	25	196	765	98	153
patients	1387	1184	216	506	954	0	116	853	201	89
children	40	18	30	25	0	0	0	1	10	13
risk	2961	2496	1220	98	201	10	44	116	500	37
effects	452	260	262	459	89	5	87	556	65	40
chemotherapy	925	671	382	765	853	1	112	642	116	20
therapy	1007	857	370	422	399	0	56	351	164	16
side	324	111	186	434	62	6	80	464	50	29
time	249	288	182	196	116	0	132	112	44	11
years	674	665	956	226	313	10	41	106	191	24
feel	66	68	82	157	12	9	29	84	0	10
people	173	42	33	85	27	7	9	147	21	33
life	200	212	223	153	89	13	11	20	37	126
family	356	345	121	5	34	8	0	2	110	6
radiation	295	148	87	303	184	13	33	190	73	20
child	24	27	31	81	0	1	7	8	5	0
cells	1776	756	167	303	35	0	31	311	57	0
supportive										
cancer	2962	2984	551	1196	850	1115	529	185	203	734
breast	2984	686	449	369	138	113	129	116	136	136
women	551	449	334	139	31	53	107	138	20	68
treatment	1196	369	139	828	247	175	253	114	38	197
patients	850	138	31	247	418	20	100	70	26	166
children	1115	113	53	175	20	2278	511	39	4	361
risk	203	136	20	38	26	4	24	8	84	43
effects	226	92	74	444	93	41	68	76	0	13
chemotherapy	185	116	138	114	70	39	49	116	8	38
therapy	150	107	70	173	116	25	28	134	24	17
side	167	133	61	408	63	20	56	60	21	19
time	529	129	107	253	100	511	452	49	24	233
years	368	116	45	118	83	219	78	19	38	123
feel	643	167	233	252	273	522	274	67	14	245
people	669	47	40	131	123	438	296	21	40	208
life	734	136	68	197	166	361	233	38	43	696
family	753	103	54	117	33	577	289	0	30	333
radiation	211	263	96	532	53	6	68	174	62	5
child	288	15	1	42	12	786	249	3	0	195
cells	110	21	0	28	19	22	15	12	0	7

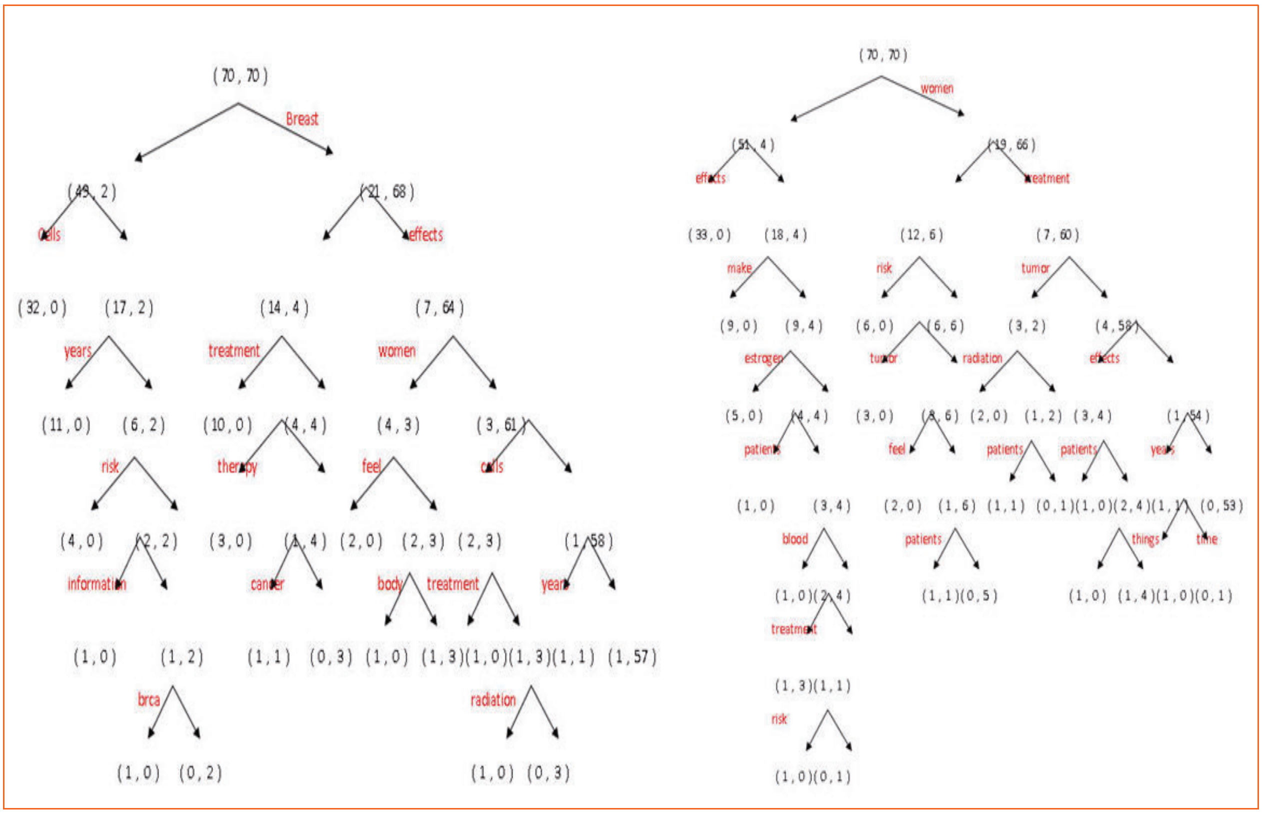

Figure 2.Examples of two decision trees for classification of Medical versus Supportive web sites (from separate folds of an 8-fold cross validation process). Each decision node is annotated with the number of cases (Supportive, Medical) and the word(in red) used to fifferentiate at that node.

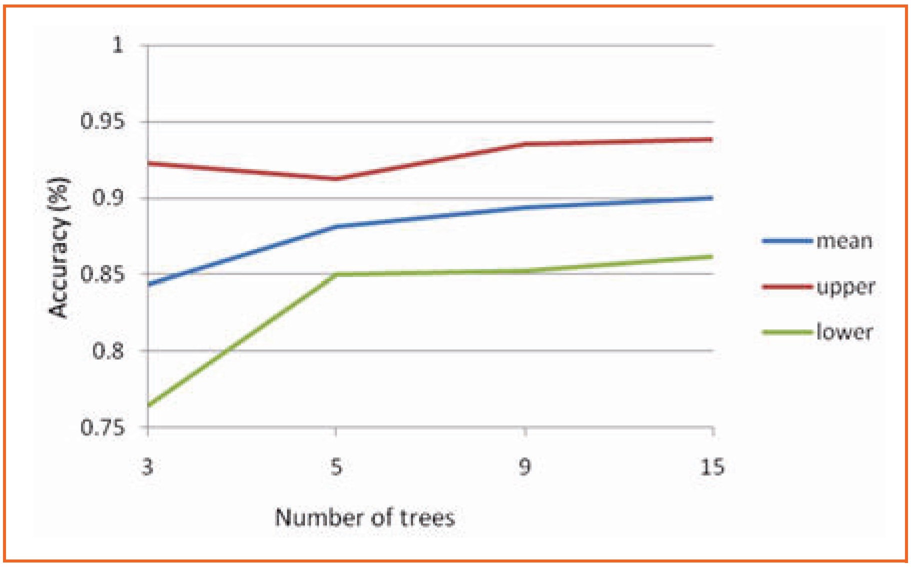

Figure 3. Classification accuracy (mean, and upper and lower bounds of 95% confidence interval) for various numbers of independent trees in a decision forest (based on 8-fold cross validation).

3. Results

Table 1 shows the HAL matrices for the Medical and Supportive samples (showing just the subset of the matrices with the 10 words with the highest HAL scores and the 20 words with the largest co-occurrence). Differences are apparent – for example, consider the larger use of the word 'children' against the 10 dominant words in Supportive versus Medical pages. With such visually-apparent differences in the HAL matrix it is unsurprising that we should be able to develop classifiers that can distinguish these differences automatically. Decision tree classifiers are developed on a larger matrix of the top 100 highest scoring words from each of the Supportive and Medical training sets.

Figure 2 illustrates two specific decision trees that emerge from the training process. The trees have good face validity with words of obvious relevance to the topic (i.e., breast cancer and consumer advice regarding it) dominating the trees and particularly central words (such as "breast" and "women") occupying the key root node positions. Note that repeat occurrence of the same word in different branches of a tree (e.g., the word "patients") is not contradictory, although multiple appearances of the same word on the same branch would be. Note that the two trees in figure 2 are from separate "folds" (divisions of training and testing data) in the 8-fold cross validation process where words are divided into 9 distinct groups to form 9 independent decision trees in a decision forest.

We attempt several forest sizes and see accuracy rising modestly as we move from 3 to 15 independent decision trees (figure 3). It can be seen that classification test results for BCKOnline Medical versus Supportive type resources are 90% accurate (95% confidence interval, 86-94%) using a decision forest classifier with 15 trees. While the variance is too large for a definite conclusion, it appears that applying the concept of a forest (using multiple trees which 'vote' of the decision outcome) provides a substantial benefit (noting the sharp rise in accuracy using five trees versus three, and the incremental but continued improvement as we move to nine and then 15 trees). The observed classification accuracy over 85% is appealing and clearly provides significant decision information for a problem where a 'coin flip' would yield only 50% accuracy.

4. Discussion

When confronted with a challenge such as a breast cancer diagnosis, health consumers will require a range of qualitatively distinct types of information, including information on local resources and humanizing perspectives. This work aims to facilitate classification of types of consumer web source resources. Early results are indicating that language use patterns can be used to automate such classification with significant accuracy. Classification results of around 90% accuracy observed for the two-case classification problem of distinguishing Supportive versus Medical tone breast cancer consumer information sites when using manually pre-classified websites as training data.

The current results still need to be regarded with some caution. A wider range of websites and metadata attributes needs to be assessed and compared to end-user feedback. However, visual inspection of the HAL matrices (see Table 1), indicates that the success of the classifier is unsurprising – there are differences in language use that can be quantified in terms of word co-occurrence rates. Moreover, we have demonstrated that either k-nearest neighbour or decision tree based methods can be employed to arrive at the desired classifications from the HAL matrix features. Nonetheless, the limited and specialised scope of the corpus applied to date should be taken as a limitation of the present research; it is not clear how broadly these methods can be applied.

One future application of the automated classifiers may be in a tool to facilitate metadata coders in populating the databases of domain-specific portals such as BCKOnline. Another possible application (which will be more demanding on the processing speed of the algorithm) would be in providing tagging or sorting on content type on live search results from health consumers. In either case, 90% accuracy should be adequate to provide worthwhile benefits to users.

5. Conclusion

We have trained a classifier to obtain 90% accuracy for the two-case classification problem of distinguishing Supportive versus Medical tone breast cancer consumer information sites when using manually pre-classified websites as training data. Such classification is underpinned by the features derived from a Hyperspace Analogue to Language (HAL) matrix based on word co-occurrence patterns. This research must be confirmed on a wider range of websites and meta-data attributes. The results are promising for provision of support tools for metadata coders as well as health consumers.

Acknowledgements

This research is supported in part by a University of Auckland Postgraduate Research Scholarship and ARC Discovery Project DP066535.

References

1. Madden M., Fox S. Finding Answers online in sickness and in health. Pew Internet & American Life Project, 2 May 2006. Available from: http://www.pewinternet.org/PPF/r/183/report_display.asp.

2. Kyunghye K, Mia L, Lustria D, Nahyun K. Predictors of cancer information overload: findings from a national survey. Information Research 2007; 12(4). Available from: http://InformationR.net/ir/12-4/paper326.html.

3. Eysenbach G, Powell J, Kuss O, Sa E. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. JAMA 2002; 287(20):2691-700.

4. Haynes R, Cotoi C, Holland J, Walters L, Wilczynski N, Jedraszewski D, McKinlay, et al. Second-order peer review of the medical literature for clinical practitioners. JAMA 2006; 295(15): 1801-8.

5. Moon J, Burstein F. Intelligent portals for supporting medical information needs. In Arthur Tatnall (ed.), Web Portals: The New Gateways to Internet Information and Services, Idea Group Publishing: Hershey Pennsylvania, 2005, pp. 270-96.

6. Madden A. Portals or filters? Identifying quality on the internet. In Andrew Cox (ed.), Portals - People, Processes and Technology, Facet Publishing: London, 2006, pp. 14-23.Burstein F, Fisher J, McKemmish S, Manaszewicz R, Malhotra P. User centric portal design for quality health information provision. Proce, 38th Annual Hawaii International Conference on System Sciences, IEEE Computer Society: Los Alamitos, California, 2005.

7. Enterprise Information Research Group, Monash University. Breast Cancer Knowledge Online Portal, with BreastCare Victoria and the Breast Cancer Action Group, Monash University, 2004. Available from: http://www.bckonline.monash.edu.au/.

8. Hunter J. Next generation tools and services: supporting dynamic knowledge spaces. In Cushla Kapitzke and Betram C. Bruce (eds), Libr@ries: Changing Information Space and Practice, Lawrence Erlbaum Associates: Mahwah, New Jersey, pp. 91-111, 2006.

9. Burgess C, Lund K. Modelling parsing constraints with high-dimensional context space. Language and Cognitive Processes 1997; 12(2/3): 177-210.

10. Burgess C, Lund K. Representing abstract words and emotional connotation in a high-dimensional memory space. Cognitive Science Proceedings 1997, LEA. pp. 61-66. Available from: http://hal.ucr.edu/pdfs/Burgess_Lund(1997b).pdf.

11. Chen J, Warren J, Yang T, Kecman V. Adaptive k-local hyperplane (AKLH) classifiers on semantic spaces to determine health consumer webpage metadata. In Proc, 21st IEEE International Symposium on Computer-Based Medical Systems, Jyväskylä, Finland, June 2008, pp. 287-9.

12. Yang T, Kecman V. Adaptive local hyperplane classification. Neurocomputing 2008; 79(13-15):3001-4.

13. Chen G, Warren J, Evans J. Automatically generated consumer health metadata using semantic spaces. Conferences in Research and Practice in Information Technology 2008;. 80 (Health Data and Knowledge Management): 3-9, Wollongong, January; also Australian Computer Science Communications; 30(9).

14. Ho T. The random subspace method for constructing decision forests. IEEE Trans. On Pattern Analysis and Machine Intelligence 1998; 20(8): 832-44.

Correspondence

Prof. Jim Warren
Department of Computer Science - Tamaki
The University of Auckland
Private Bag 92019
Auckland 1142, New Zealand

Phone: +64 (0)9 3737 599
Fax: +64 (0)9 3737 503
http://www.cs.auckland.ac.nz/~jim/

jim@cs.auckland.ac.nz