Title	WWW上のがん情報の分類に関する研究
Author(s)	木村，俊也
Citation	
Issue Date	2007-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/3598
Rights	
Description	Supervisor:島津　明，情報科学研究科，修士

Automatic Classification of Cancer Information on WWW

Shunya Kimura (510030)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 8, 2007

Keywords: text classification, web mining, machin learning.

By the development of recent Internet Technology, it is possible to offer various information about cancer via the Web. The Web is becoming a kind of important information infrastructure for the patients and their families. In this study, we note information about cancer whose demand is high. The reason of growth of cancer information is that there are a lot of medical treatments about cancer. Cancer Patients and their families search new and good medical treatment to get better medical treatment about cancer.

To get the newest information about cancer, we think it is a “medicine” like a “operative treatment” and “ internal medicine.” From such a background, we investigated cancer information on Web pages with a medical doctor, and pointed problems. It is difficult for usual people to select appropriate information by using general search engines. If we search WWW for cancer information with a general search engine, we get a various information about cancer such as information written by medical doctors, information written by cancer patients (cancer diary) and commercial information. Commercial information occasionally confuses cancer patients because it mixes information explaining cancer disease and advertisement about cancer medicine. The purpose of this study is to clarify a method which supports people who have no expert knowledge about cancer so that they can select cancer information correctly via Web.

We found that cancer information on the Web can be classified into five categories about five kinds of Cancers (Stomach cancer, Lung cancer, Colon cancer, Uterine cancer, Leukemia). We call these 5 categories CII (Cancer Information Index) as follows: (C1) Authorized information by special medical institutions (reliable but difficult for patients to understand), (C2) Personal information (close to patients’ needs) (C3) Media information (portal sites and information about book) (C4) Commercial information 5. Noise (Web pages without cancer information). Advanced information in the authorized information category includes useful information but is difficult to understand for patients having no knowledge about cancer. We think information in personal information category made by individuals is the closest to patients’ needs.

For the problem that information obtained by a search engine is diverse and disordered, we think that if we offer patients information classified according to CII, patients come to be able to select dependable information. Accordingly, to solve the problem, Nakagawa made dictionary including technical terms about cancer (3316 words) since computers cannot recognize technical terms about cancer. The dictionary was made from sentences of Web site “National Cancer Center” by hand.

We made a Naive Bayesian classifier that classifies cancer information automatically according to CII, and experimented classification test. The accuracy of the classification is about 80% and we figure out that this classifier has enough performance to classify cancer information. However, from the discussion about the classification result, we found out that it is difficult to classify C4 category with only language features. We guess reasons of this problem as follows: (1) C4 includes information which maliciously leads patients to purchase goods. Some parts of such pages have information for sales and the other parts explain cancer. These confuse the estimation for the correct category. (2) Pages of individuals or some companies refer Web pages of C1 to explain cancer diseases. Considering these, we use web morphological features such as image size and html size of web pages in addition to language features to classify web pages correctly. Accordingly, the classification accuracy is improved.