WWW 上のがん情報分類に関する研究

木村 俊也 (510030)

北陸先端科学技術大学院大学 情報科学研究科

2007 年 2 月 8 日

キーワード： 文書分類，ウェブマイニング，機械学習.

昨今インターネット技術が発達し，ウェブを介してさまざまな情報提供が行われるようになってきており，ウェブ上の医療に関する情報が日々増加している．医療患者やその家族にとってウェブは重要な情報基盤のひとつになりつつある．本研究では医療情報の中でも需要の高いがん (癌) 情報に注目して研究している．がん情報が他の医療情報に比べて盛んに流通するのは，治療法が確立されつつある糖尿病や循環器疾患に比べ，施設間での診断・治療に関する見解が標準化されておらず，診断治療にあたる医師や医療機関によって生存率が異なることが問題となっているなどの背景がある．がんを宣告された患者や家族は新しく可能性のある治療法を検索し治癒の可能性の高い医療機関に移りたいという要求から少しでも多くの情報を得ようとする．

我々は，最新のがん情報を的確に得ることは延命や治療のために，手術，内服薬に匹敵する第 3 の薬であると考えている．このような背景のもとで，ウェブ上のがん情報に関する調査を医師とともに行い，以下のような問題点があることを報告した．検索エンジンを用いてがん情報を検索すると，医師が記述したものや個人が記述したもの（闘病記など），商用の情報などが無秩序に出力され，医学に関する専門的な知識を持たない一般人にとってはどの情報が正しいのかの判断が困難である可能性が高いことを指摘した．商用のがん情報ページには，有用でありうるがんに関する情報が記述されているが，商用誘導を企てているページが存在するため，がんの治療法を探しているがん患者を困惑させてしまう可能性が高い．以上の問題を解決し，がんに関する専門知識がない一般人にも，がんの情報を正しく選択できるように支援をすることが本研究の目的である．

胃がん，肺がん，大腸がん，子宮がん，白血病の５つのがんについて，わが国で発信されているこの分野のコンテンツは次の 5 類型に分類できることを示した．この 5 類型を我々は CII(Cancer Information Index) と呼び，次のような類型となっている．1. 専門医療機関などの高度な内容（有用性は高いが患者には理解が困難な内容のものがある），2. 個人医師や患者個人による患者志向の内容（患者のニーズに近い情報が多い），3. 個人を対象としたポータルサイトや書籍情報，4. 個人を対象とした商用情報，5. がんの情報を含まないもの．このうち専門性の高い研究志向の情報である類型１は根拠があり有用な情報

を含むが，専門用語の知識のない患者にとって理解することが困難であり，間違った解釈を生むことも考えられる．類型２の個人を対象とした個人（情報ボランティア）による情報発信からの情報がより患者のニーズに近い情報を与える可能性を示唆した．

そこで我々は先に挙げた問題である「様々な情報が無秩序に提供される問題」を解決するために，CII に従った自動分類を行った情報を患者に提供すれば患者は,「信頼できる情報」を選別することが可能であると考えた．これを実現させるために，まずがんの情報で特徴的に使用される専門用語を計算機が認識できないという問題を解決するため，中川ら [20] が作成したがんに関する専門用語辞書３３１６語を用いることにした．がん用語辞書は国立のがんの専門研究機関である国立がんセンターのウェブページにある，がん疾患解説ページから手作業で作成した．

次に，がん情報の分類の予備実験としてがんに関するウェブページを機械学習の手法を用いて自動的に CII に従って分類する分類器を作成した．分類実験をした結果，分類精度は約 80%が得られ，自動分類に使うには十分な成果を得られた．しかし，分類実験の考察で先に挙げた分類実験で類型 4（商品情報などの）のウェブページは言語情報だけでは自動分類が困難であることが示唆された．この問題は次の２つが主な原因であると考えられる．1.C-4 には商用誘導を企むページが存在し，ウェブページ上に販売を目的とした箇所と，がんの疾患を解説するための箇所が混在しているページが存在する．2. 個人や業者ががんの疾患を説明するために公的な機関によって発信されたウェブページを参照して記述したページが存在する．上記の 2 点の問題を解決するために言語情報に加えて，イメージの数や，ファイルの総量といったウェブの形態的な情報を用いて分類する手法を提案した．この手法により，言語情報だけで分類するよりも分類精度が向上することを示した．