商品検索を対象とした相関計量を用いた対話型問い合わせ自動生成システム

百々 健人† 清木 康‡

†慶應義塾大学環境情報学部 〒252-0805 神奈川県藤沢市遠藤 5322

E-mail: †t12552kd@sfc.keio.ac.jp, ‡kiyoki@sfc.keio.ac.jp

あらまし 本稿では、商品情報を構成している文書に対してテキスト解析を行いメタデータに変換し、多次元空間上で相関計量を行うことで、検索のための問い合わせを自動生成するシステムを提案する。

本方式は、商品の説明文を解析し、特定の条件に適合する単語のうち出現頻度の高い単語を主軸として商品に対する多次元空間を作成することで、相関計量によって商品の意味を空間上に定義する。その後に、作成した多次元空間の主軸から適切に問い合わせを自動生成する。本方式の実現により、既存では人力で作成している条件分岐による問診型検索を自動生成でき、かつ広い範囲のカテゴリーの商品を対象とした問診型検索の生成が可能となるので、より幅広い商品情報へのアクセスが可能となる。

キーワード 商品検索, テキストマイニング, 感性メタデータ, 多次元空間計量, 問い合わせ自動生成

1. はじめに

近年、インターネットと物流の発達により、web 上で商品を購入する人が増え、2014 年度の日本国内における BtoC の EC 市場物販系分野の規模は、6 兆 8043 億円に達し、なおも成長を続けている[1]。

しかしながら、未だ物販の EC 化は 5%未満である。言い換えれば、消費者は残りの約 95%のほとんどは実店舗で商品を購入するのである。

物販の EC 化が急速に進まないことについて、物販系の EC サイトのうち多くを占める商品を検索し購入することが可能なサイト（以下、商品検索サイトとする）は、実店舗に比べて、感性（未知の単語）に関する検索に弱いことが原因の一つと考える。医薬品の例を挙げると、多くの商品検索サイトでは文字列検索の他にカテゴリー検索が可能であるが、医薬品を購入したい人が感じる症状ごとに分かれていないことが多い。また、仮に分かれていたとしても、最も強く感じる症状やその原因との関係は年齢・性別等により異なる。よって、問題適応型計算を取り込むことが可能で網羅性のある検索方式を確立する必要がある。

そこで活用できるのが、対話型検索である。この方式は、あらかじめ質問を設定し、その質問に対する回答によってクエリを生成して検索をして結果を表示するというものである。ところが、対話型検索は、質問やクエリ生成規則を作成するのに人間の労力が必要となるデータ数が非常に多い商品データにこの方式をそのまま適用することは難しい。よって、本研究では、問題適用型計算を取り込むことが可能な網羅性のある商品検索を対象とした対話型問い合わせを自動生成するシステムを提案する。

2. 商品検索を対象とした相関計量を用いた対話型問い合わせ自動生成システムの方式

本システムの基礎的な方式については、先行研究[2]で述べたものを応用したものである。よって、本稿では、本研究の新規発展部分を含まない部分は割愛する。

2.1. システムの全体構造

本システムは、3 つの部分に分割することができる。3 つの部分とは、商品情報データベースから商品の説明文を取り出し文章解析処理を行うテキスト解析部、解析したテキストを重要語となるタグを抽出しするタグ抽出部、それらタグを組み合わせて質問を作り適切な順番でユーザーに表示し検索したのちに結果出力を行う結果出力部、である。（図 1）

なお、3 つの部分とは別に、人間が認知できない現象を取り除く質問選定のプロセスが存在する。

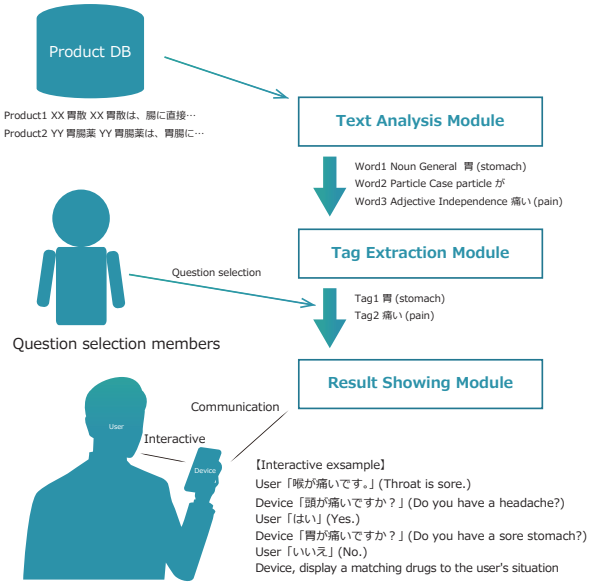


図 1 システムの全体構造と具体例

2.2. システム基本方式

本システムを構成するにあたって必要となる各部分の方式について記述する。

2.2.1. 文章解析方式

文章解析部では、形態素解析を用いて商品の説明文を単語ごとに分割し、句読点など以後の計算で不要な単語を除外した単語データをテーブルに保存する。

2.2.2. タグ抽出計算方式

タグ抽出部では、文章解析部にて保存した単語データテーブルを使用する。単語データのうち最終的な質問文を構成する品詞・品詞細分類（具体例は 3 項で後述）に対して出現率が一定以上のものをタグとして保存する。また、タグと商品の関連度とタグ同士の関連度を次項で記述する意味の数学モデル[3][4]を用いた関連度計算モデル（2.2.3 項）を用いて算出する。

2.2.3. 意味の数学モデルを用いた関連度計算モデル

タグ抽出部で行う関連度計算のモデル（図 2）には、意味の数学モデルを用いている。

意味の数学モデルとは、共通の意味を持たない独立したメタデータを次元の軸として仮想空間を形成し、この空間上に計量するオブジェクトを定義し目的に合わせた多次元計量を行うものである。

本モデルでは、意味の数学モデルについて、質問文の一部を構成するタグからなるメタデータを次元軸として商品検索空間を形成し、商品同士とタグ同士の相関計量によって商品を空間上にプロットする。

しかし、従来の意味の数学モデルにおいて、次元の軸は独立した意味を持つ必要がある。つまり、タグをそのまま次元の軸とすると、タグ同士の関連度は 0 である（べき）はずである。ここで、空間上の軸とタグは同じ言葉であるが違う意味とする。違う意味として定義することが可能な理由は、その”感性の主語”を変えているということにある。例えば、「目が重いですか？」の「目が重い」は、実際に「目が重い」だけではなく「頭が痛い」ために感じることがある。これを言い換えると、正しくは（神経は）「目が重い」であるが、人は「頭がいたい」と感じる。ということが挙げられる。

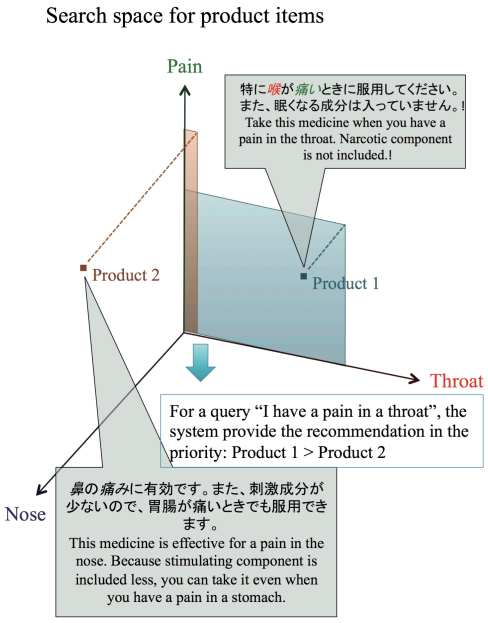


図 2 商品対話型検索を対象とした多次元意味空間

2.2.4. 対話型検索方式

結果出力部では、抽出されたタグと商品検索空間を元に質問を表示し、ユーザーのその質問に対する回答によって検索クエリを発行し、検索結果を表示する。質問の回答から検索クエリを発行し検索結果を得る計算方式については、先行研究と同様である。なお、検索結果については、検索クエリと商品の関連度が高い順に表示するものとする。

2.3. システム実現方式

本システムを実現するにあたり、データベースは MySQL を使用し、対話型検索を行う部分については、node.js を利用して HTML ページを生成、後で触れる実

験では PHP を利用した HTML ページを生成した。なお、形態素解析ツールは先行研究に引き続き MeCab[5]を使用した。

2.4. システムの特徴

本システムには以下 3 点の特徴がある。

（1）検索対象の範囲が、エキスパート検索システムに比べて広い。

（2）人力で作成している条件分岐による問診型検索を自動生成できる。

（3）既存の文字列マッチング検索よりも対話的である。

3. 検証実験と考察

本システムによる検索が有効であるかを調査するために、検証実験とその考察を行った。

3.1. 商品関連度検証実験について

先行研究では、関連度計算モデルの確からしさを検証するため、商品関連度検証実験を行った。この実験では、医薬品 500 品目からランダムに商品を選択したのちに、その商品と関連がある商品を正解とし、一方で商品検索空間から商品の関連度を求め、閾値以上の関連度を持った商品リストと正解リストについて、再現率・適合率・尺度 F 値[6]を求めた。

商品関連度検証実験では、尺度 F 値について、閾値 0.3 において 0.15～0.43 と不安定な結果となった。この原因として、特定の商品に対して関連のある商品の選択自体の難易度が高かったことでそれがノイズの原因になったと考えられる。

3.2. 検索結果の確からしさ検証実験

次に、本システムに対して実際の検索結果を対象として実験を行った。

3.2.1. 実験手法

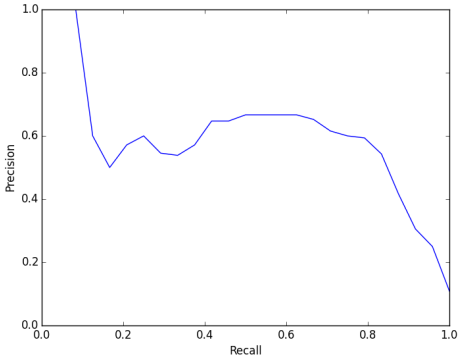
本実験は、医薬品通販サイトであるケンコーコム[7]上で取り扱われている医薬品 5420 品目を対象にタグ抽出部終端に至る問い合わせ自動生成計算を行い、質問候補を取り出せる状態にした上で、質問候補のうち全商品説明文で出現率が上位のものについて検索精度の検証を行った。ここで、より具体的には、上位 3 つの質問それぞれ「はい」と答えたと仮定して、それぞれにランダムで抽出した 500 品目のうちどの商品が当てはまるかを選択し正解リストを作る。その上で、それぞれの検索結果について変数となる閾値を上回る商品リストとの再現率・適合率・尺度 F 値を求め、検証した。

3.2.2. 実験結果

各質問のうち、最大の尺度 F 値が最大と最小になる質問に対する再現率・適合率・尺度 F 値、Recall-Precision グラフ[8]は以下の通りとなった。

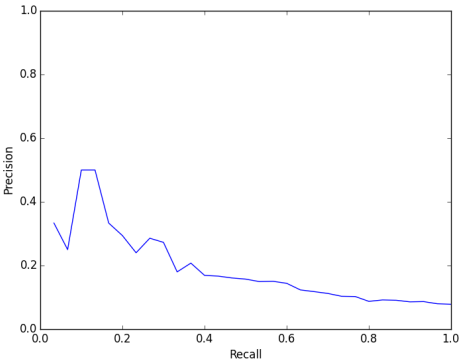
● 胃が痛い

閾値	検索結果数	検索結果中正解数	正解数	再現率	適合率	尺度F値
0.1	120	23	24	0.958	0.192	0.320
0.15	51	21	24	0.875	0.412	0.560
0.2	30	18	24	0.750	0.600	0.667
0.25	17	11	24	0.458	0.647	0.536
0.3	13	7	24	0.292	0.538	0.379



● 体がだるい

閾値	検索結果数	検索結果中正解数	正解数	適合率	再現率	尺度F値
0.1	118	17	30	0.144	0.567	0.230
0.15	45	9	30	0.200	0.300	0.240
0.2	18	6	30	0.333	0.200	0.250
0.3	6	3	30	0.500	0.100	0.167



「胃が痛い」の検索結果については、閾値 0.2～0.25 を中心に適合率･再現率ともに高い数値となっている。一方、「身体がだるい」の検索結果については尺度 F 値が最大 0.250 である。よって、本システムは検索精

度が最低限確保されていると考えられる。

4. システムの応用性

本システムは、説明文をメタデータに変換した情報のみ使用することから、日本語以外の言語で記述された説明文に対しても同じ方式を適用できる。

また、検索空間モデル上での相関計量が検索の結果に反映されるため、パーソナル情報を利用して空間モデル上の軸に重み付けを行い年齢・性別によって結果を変えることが容易に可能である。

参　考　文　献

[1] 平成 26 年度我が国経済社会の情報化・サービス化に係る基盤整備（電子商取引に関する市場調査）, p.2.

[2] 百々 健人, 清木 康, “対話型商品検索レコメンドを対象とした問い合わせ自動生成システム”, DEIM2015, pp.1-pp.5, 2015.

[3] T.Kitagawa and Y.Kiyoki, “A mathematical Model of Meaning and its Application to Multidatabase Systems,” Proceedings of 3rd Issues on Data Engineering: Interoperability in Multidatabase Systems, pp.130-135, April 1993.

[4] 中村 恭子, 金子 昌史, 清木 康, 北川 高嗣 “意味の数学モデルによる意味的画像探索方式とその学習機構”, Information Processing Society of Japan, 1995.

[5] MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.

[6] D.M. Christopher, R. Prabhakar, and S. Hinrich: Introduction to Information Retrieval, Cambridge University Press, New York, 2008.

[7] ケンコーコム: http://www.kenko.com/

[8] “The Relationship between Recall and Precision”, JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, January 1994