Title	対話型レコメンド検索を対象とした問い合わせ自動生成システム
Sub Title	
Author	百々, 健人(Dodo, Kento) 清木, 康(Kiyoki, Yasushi)
Publisher	慶應義塾大学湘南藤沢学会
Publication year	2014
Jtitle	交通運輸情報プロジェクトレビュー No.23 (2014.) ,p.56- 60
JaLC DOI	
Abstract	本稿は, オブジェクト(対象物)に紐づいている文書情報をテキストマイニングして量的データに変換し, 多次元空間計量を行うことで, 問い合わせ自体とその結果を自動生成するシステムを提案する。 本方式では, オブジェクトの一例として商品情報を取り扱う。本方式は, 商品情報の一部である説明文を解析し, 出現数を調べる。条件に合致しかつ出現数が高いものは, 重要な単語としてそれら同士の関連度を計算する。その後に, これらの情報から質問を自動生成する。これにより, 既存では人力で作成している条件分岐による対話型検索を自動生成でき, 労力を低減することを目指す。
Notes	2014年度慶應義塾大学JR東日本寄附講座報告書 慶應義塾大学交通運輸情報プロジェクト その3 : 慶應義塾SFC大学院生•学部生の研究
Genre	Technical Report
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO92001006-00000023-0056

対話型レコメンド検索を対象とした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 本稿は、オブジェクト（対象物）に紐づいている文書情報をテキストマイニングして量的データに変換し、多次元空間計量を行うことで、問い合わせ自体とその結果を自動生成するシステムを提案する。

本方式では、オブジェクトの一例として商品情報を取り扱う。本方式は、商品情報の一部である説明文を解析し、出現数を調べる。条件に合致しかつ出現数が高いものは、重要な単語としてそれら同士の関連度を計算する。その後に、これらの情報から質問を自動生成する。これにより、既存では人力で作成している条件分岐による対話型検索を自動生成でき、労力を低減することを目指す。

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

インターネットが一般に普及し発達してきたことで、誰もがweb上で商品情報を検索し購入することが可能になりつつある。さらには、流通の発達により数日もしないうちに商品が届けられるようになったことで、その便利さ故に商品情報を検索し購入できるサイト（以下、商品検索サイトとする）の利用者は増加の一途を辿っている。

しかしながら、既存の商品検索サイトには便利さを向上する余地がある。特に、商品検索の部分に関しては、現状では商品カテゴリーの分類や価格帯といった設定されたグループに絞り込む、もしくは文字列で検索する、またはその両方である。確かに、この検索手法であれば、具体的に欲しいものが想像できるときには有効である。ところが、商品検索サイトのユーザー（以下、ユーザーとする）自身が具体的に欲しいものが想像できない状態の場合は、必ずしも潜在的な欲求に応じることはできないと考えられる。

そこで活用できるのが、対話型の検索（以下、対話型検索とする）である。これは、あらかじめ設定した質問に対する回答によって検索結果を表示するというものである。この検索手法を商品検索に取り入れることによって、ユーザーの潜在的欲求を満たすことができると考える。

ところが、対話型検索は、質問とそれを表示するフローやその回答に対する検索結果を作成するのに人間の力が必要となり、データ数が非常に多い商品データにこれをそのまま適用することは難しい。よって、本研究では、省労力でユーザーの潜在的欲求（特に、感性から表現する欲求）を満たせるよう、対話型商品検索を対象として問い合わせを自動生成するシステムを提案する。

2. 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の手法

2.1. システム構造

本システムは、商品情報の説明文を文章解析する部分、その情報を元に検索時に出題する質問を構成する単語（質問群を構成するためのメタデータとなる。これをタグと呼ぶこととする）を選定し抽出する部分、それらの情報より実際に対話型検索を行い、表示する部分の３つで構成される。（図１）

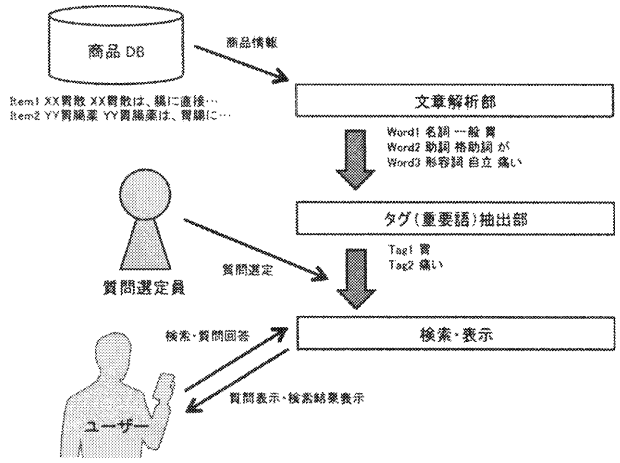


図１ システムの基本構造

まず、商品データベースから商品情報を取り出し、商品の説明文に対して文章解析を行う。次に、文章解析の結果をもとにタグを抽出する。タグは、最終的に質問を構成するために必要な品詞・品詞細分類の組み合わせの条件とある一定上の重要度を担保するために出現率の条件によって、抽出する。これによって、実際に使用する質問文群の候補が出来上がる。この後、質問選定員によって、質問文を最終選定し、質問文群

ができあがる。これら質問文をユーザーが最初に選択もしくは入力したクエリに応じて表示し回答を受け付けることで、商品検索に必要な情報を収集し、検索結果を表示する。

2.2. 本システムの特徴

本システムの特徴は、（1）検索結果表示までに最低限必要な情報は、商品の名前と説明文のみである。（2）検索結果表示までに必要な人間の労力は、商品を選ぶときに重要な質問群をその候補から選択するのみである。（3）メタデータを用いての対話型検索を行える。ことである。

2.3. 基本方式

本システムを構成するにあたって必要となる計算等の方式について、詳細を記述する。

2.3.1. 文章解析方式

文章解析部では、主に形態素解析を用いる。商品の説明文に対して形態素解析を行い単語ごとに分割し、句読点やスペースといった計算に不要な部分（stop words）を除外した単語データを保存する。

2.3.2. タグ（重要語）抽出計算方式

タグ抽出部では、タグを抽出するために文章解析部にて保存した解析済みの単語データを用いる。単語データのうち最終的な質問文を構成する品詞・品詞細分類のみを選択し、それぞれの品詞・品詞細分類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、同時にそれらタグと商品の関連度とタグ同士の関連度を、意味の数学モデル[1][2]を主モデルとして算出を行う。意味の数学モデルを用いた商品とタグの関連度計算方式については、次項で記述する。

2.3.3. 意味の数学モデルを用いた関連度計算方式

意味の数学モデルは、個々に異なる意味を持ったメタデータを次元の軸として空間を形成するモデルである。このモデルを、商品対話型検索を対象とした多次元意味空間の形成に活用する。具体的には、タグをメタデータとしてこれを次元軸と仮定し、説明文におけるタグの出現数によって商品をプロットする。このとき、商品とメタデータの関連度が決定する。（図2）

一方で、メタデータとしてのタグはこの時点では最終的な質問文中に出現する単語のそれとは意味が違うという点がある。これは、タグについて単語として登場するときと文中に登場するときは意味が必ずしも一致しないということから来ている。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」ために感じることもあるからである。そのような差異を是正するため、あらかじめ次元軸上のメタデータと質問としてのタグ（実質、タグ同士）の関連度を求めておく。

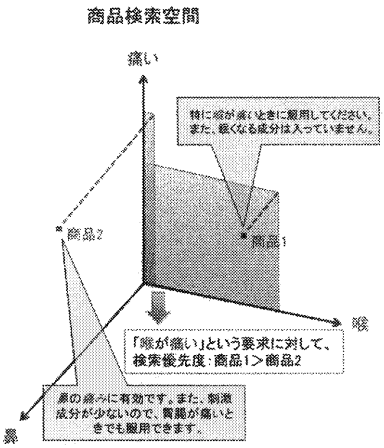


図2 商品対話型検索を対象とした多次元意味空間

2.3.4. 質問文選定方式

タグ抽出部でタグを抽出した後に、タグ群から質問文群を選定する。

まず、全商品の説明文中のうち最終的な質問文を構成する品詞・品詞細分類に当てはまるもののうち、タグで構成されているもの（ただし、助詞などの単体で意味を持たない品詞の単語は除く）を質問文群の候補とする。（これを、質問文群候補抽出とする。）

次に、この群より人が商品を選択する上で重要な質問となる質問を選定する。これは実際に人が候補からの選定を行うが、選定基準として、質問文候補に対し（1）助詞の変更と語尾の追加・変更により意味が通るものである（2）最終的に生成される質問文の意味が曖昧でないこと（3）それが短時間（数秒単位）で定量的に変換できる回答が可能な質問であること。の3点を設定し、選定を行った。（これを、質問文選定とする。）

2.3.5. 対話型検索方式

いくつか出題する質問文の回答を元に、商品検索を行う。商品検索時に必要な情報はそれぞれの商品とランダムに表示される質問文の回答との関連度（これを、感性関連度とする）である。この大きさで、検索候補と出現準を定める。

ここで、当システムにおいて、感性関連度の算出式を以下のように定めた。

ある商品Iに対しての感性関連度をV、単一質問Q_nに対する感性関連度をV_{Q_n}と定義すると、

$$V = \Sigma_n V_{Q_n}$$

一方で、質問Q_nはタグ群$(Q_nT_1, Q_nT_2, \cdots, Q_nT_h)$から構成していることから質問はタグ群の和集合と考えられるため、質問とタグQ_nT_mとの関連度を$V_{Q_nT_m}$と定義すると、

$$V_{Q_n} = V_{Q_nT_1} \times V_{Q_nT_2} \times \cdots \times V_{Q_nT_h}$$

ここで、商品Iが保持するタグ（説明文内に出現するタグ、重複タグは削除しない）とQ_nT_mと関連のある（関連度が0より大きい）タグの集合を$(Q_nT_mR_1, Q_nT_mR_2, \cdots, Q_nT_mR_j)$と定義し、それらの関連度を$(V_{Q_nT_mR_1}, V_{Q_nT_mR_2}, \cdots, V_{Q_nT_mR_j})$とすると、

$$V_{Q_nT_m} = \Sigma_x V_{Q_nT_mR_x}$$

$V_{Q_nT_mR_x}$に関しては、タグ抽出部で算出した値を使用する。

以上を元に、質問の回答から商品検索の結果を得ることとした。

3. 実現方式

本システムを実現するにあたって具体的に使用した事項について詳細を記述する。

3.1. プログラムと制御と表示の方式

本システムを実現するにあたって、各プログラムやその制御、HTMLページへの表示に関して、プログラミング言語のPHPを使用した。これを使用することにより、簡易的に動的なHTMLページを生成し表示することが可能である。

3.2. データベースの仕様と設計

本システムを実現するにあたって、重要な役割を担うデータベースは、MySQLを選択した。本項では、システム内の各部において使用するデータベーステーブルについて、記述する。

3.2.1. 商品情報テーブル

本システムの対象物である商品情報を格納するテーブルでは、1つの商品を1つのタプルをしてデータを保存する。以下に、テーブルの構造とシステム動作に最低限必要な情報かどうかを示す（図3）。

カラム名	型	主キー	必要な情報
iID	int	○	○
name	text		○
description	text		○
image	text		×
maker	text		×
brand	text		×

図3 商品テーブルについて

システム動作に最低限必要な情報が全て存在する商品が本システムにおける検索対象となる。

3.2.2. 単語テーブル

文章解析部では、商品の説明文を形態素解析した結果を単語ごとに単語テーブルに保存する。形態素解析を実行するにあたって、オープンソースであるMeCab[3]を用いた。

以下に、テーブルの構造とシステム動作に最低限必要な情報かどうかを示す（図4）。

カラム名	型	主キー	必要な情報
iID	int	○	○
start	int	○	○
word	text		○
psub1	text		○
psID	int		○

図4 単語テーブルについて

単語が出現する商品と場所を保存する必要があり、かつこれはユニークであるので、これら2カラムを主キーとした。また、形態素解析した結果のうち品詞再分類の2つ目以降は計算に使用しないことから、該当するカラムを設定していない。

形態素解析するにあたって、これ以外にstop wordsを保存するためのstop wordsテーブルが存在する。stop wordsテーブルに、この後の計算に不要な単語を予め予約しておくことで、形態素解析の対象から外すことができる。

さらに、品詞に関する情報は品詞テーブルに保存しここではリレーションとして品詞idが保存される。品詞に関する情報を分離することで、品詞の異なる別言語への対応が可能である。

3.2.3. タグテーブル

タグ抽出部で保存するデータは、タグそのもののデータの他にタグ同士の関係などがある。タグテーブル（図5）、タグ同士の関連度テーブル（図6）について、以下にテーブルの構造とシステム動作に最低限必要な情報かどうかを示す。

カラム名	型	主キー	必要な情報
tID	int	○	○
name	tinytext		○

図５ タグテーブルについて

カラム名	型	主キー	必要な情報
tID1	int	○	○
tID2	int	○	○
value	float		○

図６ タグ同士の関係テーブルについて

タグテーブルの name の型を tinytext に設定した背景は、文字数が多いタグ（単語）は質問候補として想定しないということがある。

4. 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。以下に、その詳細を記述する。

4.1. 商品関連度検証実験

本システムで商品に適応した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。

4.1.1. 商品関連度検証実験の方法

商品関連度検証実験では、ある商品に関して、システムにおけるその商品と異なる商品との関連度が、予め選択したその商品と関連のある商品に対してどのような再現率・適合率が得られるかを目的とした。具体的には、(１)医薬品通販サイトであるケンコーコム[4]上で取り扱われている医薬品 500 品を、無作為に商品テーブルへ保存（２）それらから無作為に選んだ 3 品について、商品情報を表示した HTML ページから関連度の大きいと思われる商品を選択（３）それらとシステムで算出した関連度について、再現率・適合率から尺度 F 値[5][6]を算出する。

4.1.2. 商品関連度検証実験の前提条件

検証実験にあたってシステム内にて商品に対する多次元意味空間を作成する。よって、最終的な質問構造を指定する必要があるため、これについて予め想定した。

医薬品に関して考えられる有効な質問文は以下の構造となる。(図７)

	単語１	接続	単語２	語尾
品詞	名詞	助詞	形容詞	---
品詞細分類	一般	格助詞	自立	---
例１	目	が	かゆい	ですか
例２	胃	が	痛い	ですか

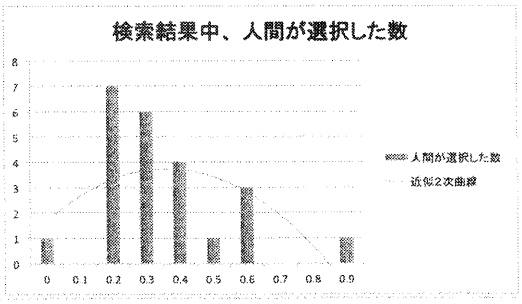
図７ 想定した有効な質問文の構造

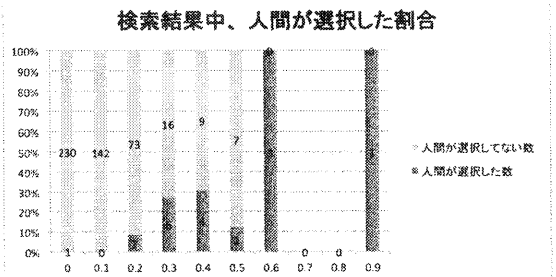
実験では、上記図中の単語１の品詞・品詞細分類をもとに商品同士の関連度を計算し、人間が選択したものと比較する。

4.1.3. 商品関連度検証実験結果

適合率および再現率を算出するにあたって、検索結果群を設定する必要がある。よって、設定に必要となる関連度の最低値（以下、閾値とする。）について適切な値を捜索するとともに、適合率・再現率とその２値の統一的尺度である尺度 F 値を求めた。

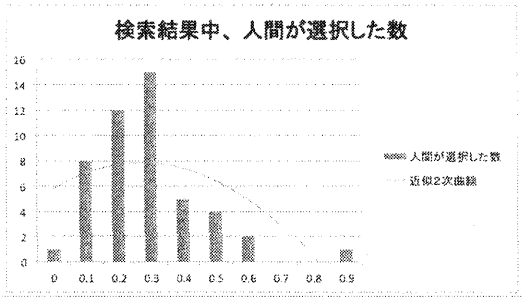
結果は、以下の図表（図表８～１０）のようになった。なお、無作為に選んだ 3 品をそれぞれ商品１、商品２、商品３とした。また、グラフ中の横軸は関連度で、小数第２位を切り捨てたときの該当数を示している。

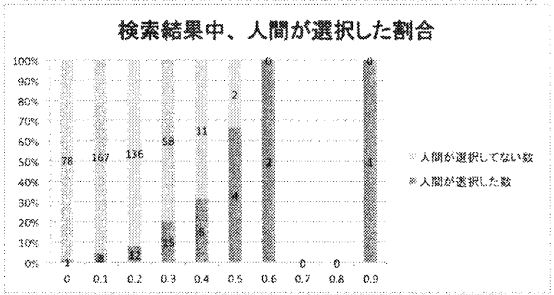




閾値	検索結果数	検索結果中正解数	正解数	適合率	再現率	尺度F値
0.2	127	22	23	0.173	0.957	0.293
0.25	78	19	23	0.244	0.826	0.376
0.3	47	15	23	0.319	0.652	0.429

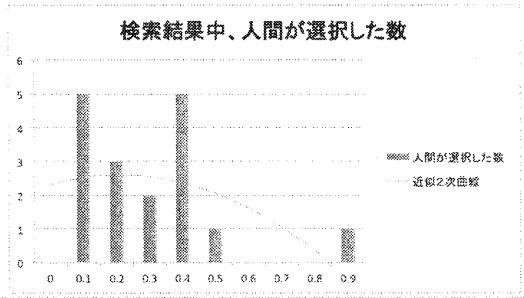
図表８ 商品１についての実験結果

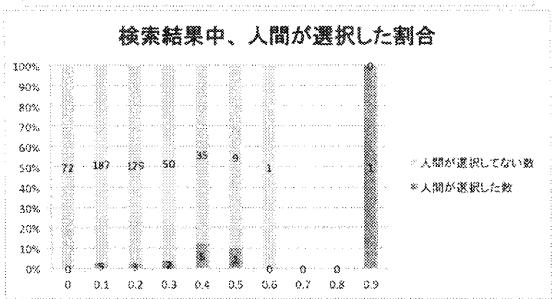




閾値	検索結果数	検索結果中正解数	正解数	適合率	再現率	尺度F値
0.2	246	39	48	0.159	0.813	0.265
0.25	151	33	48	0.219	0.688	0.332
0.3	98	27	48	0.276	0.563	0.370

図表９ 商品２についての実験結果





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
0.2	236	12	17	0.051	0.706	0.095
0.25	156	10	17	0.064	0.588	0.116
0.3	104	9	17	0.087	0.529	0.149

図表１０ 商品３についての実験結果

上記「検索結果中、人間が選択した数」について近似２次曲線の最大値が関連度 0.2～0.3 付近に存在することから、適合率・再現率を求めるにあたっての閾値について 0.2, 0.25, 0.3 を設定した。

各閾値とも再現率は比較的高いことから、検索による取りこぼしは少ないということがいえる。これは、検索商品レコメンドに適合した結果といえる。一方で適合率は低く、人間が想定していなかった商品が多く検索結果に表示されていたということになる。しかしながら、対話にて複数のクエリを作成し検索を行うので、該当しないものが検索結果より省かれる可能性が高まる。また、個々のクエリに対する検索精度については既存の検索システムを組み合わせて余分を取り除くことで解決ができる。

5. おわりに

本方式では、オブジェクトの一例として商品を扱った。

今後、本方式に適する形で対話時の質問自動選定アルゴリズムを作成し、対話型レコメンド検索自動生成システムを確立したいと考えている。これにより、駅を対象とした「駅別観光見所検索」や、事例検索・人材紹介などの説明文のついた事象についても十分応用できる。

参 考 文 献

[1] 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.

[2] 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.

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

[4] ケンコーコム: http://www.kenko.com/

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

[6] 庭野正義, マッキンケネスジェームス, 永井保夫 “適合率と再現率を用いた Web ページランキングシステムの性能評価”, 東京情報大学研究論集 Vol.14 No.1, pp.1-10(2010).