クエリとして絵文字を受け付ける情報検索

An Information Retrieval System Accepting Pictograms as a Query

山本千尋[*1] 安田宜仁[*1] 別所克人[*1] 内山俊郎[*1] 内山 匡[*1]
Chihiro Yamamoto Norihito Yasuda Katsuji Bessho Toshio Uchiyama Tadasu Uchiyama

[*1] 日本電信電話株式会社 NTTサイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

In the present retrieval system, even when a user's information need is vague, he/she must express it using several words or phrases. We propose a new information retrieval method that accepts pictograms as a part of query. In order to handle pictograms, we prepared a dictionary that maps pictograms to corresponding concepts. We implement two systems using the dictionary. One system first expands pictograms into corresponding words then conduct boolean OR-style search. The other system converts pictograms into vectors, then retrieves documents using standard vector space model. The results of these systems were compared with a conventional retrieval system.

1. はじめに

我々が普段,ある情報要求に対し検索を行う際,情報要求をもっとも適切に示しているという単語を思い浮かべ,それを検索クエリとして検索を行う. 例えば,“おいしいトマトを食べたい”という場合であれば,検索対象をもっとも適切に示す単語は“おいしいトマト”であるだろう. このように, 検索対象が明確である場合は, 検索クエリを明確にすることで,所望の検索結果を得やすくなる.

しかし,例えば映画や音楽といった検索対象の表現が漠然としている分野の検索などで,情報要求が漠然としている場合,ユーザが適切な検索クエリを言語によって表現できず,意図する内容の文書を検索することが困難な場合がある.

例えば,“明るい気分になれるような,かわいらしい映画を調べたい”という場合を考える. このような漠然とした情報要求に対し, 検索を行う場合, 言語によってクエリを表現することが難しい. しかし,従来の情報検索システムは, 文字列をクエリとして受け付けるためユーザは情報要求に言語化せざるを得ない. 前記のような情報要求を持つユーザであれば,例えば,“かわいい 映画”のような文字列をとクエリとして選択することになると考えられる. しかし,“明るい気分になれるような,かわいらしい映画を調べたい”という情報要求を,検索クエリとして言語化することによって, 情報要求に含まれる漠然とした要素,例えば,“キュート”,“明るいイメージ”などの要素が欠落した状態で検索が行われる.

漠然とした情報要求を解決する手段としてクエリ拡張技術や, ベクトル空間モデルを用いた検索技術の研究を利用することも考えられる[Voorhees1994][Qiu1993][Salton1983]. しかし,これらの手法では,ユーザは漠然とした情報要求を言語化した上で,検索を行う必要があり,ユーザの漠然とした情報要求に対して,そのままクエリを受け付けるものではない.

本研究ではユーザの漠然とした情報要求を,漠然としたままで検索するための手段として絵文字に着目した. 絵文字は,主に携帯電話によるメールや,ブログなどにおいて広く利用されている. 絵文字の特徴として,“絵”としての特徴と,文字としての特徴を持っていることがあげられる. このため,入力する文字数を大幅に減らすことができ,また,言葉で表現することが難しい漠然とした表現を表すことが可能になる. また,近年の携帯電話の入力方式を考えた場合,絵文字は日本語の入力機能に統合されている. このため,一般記号を入力するのと同じように入力することが可能であることから, “絵”でありながら,文字入力と同じように簡単に入力することが可能である.

本研究では, 絵文字の持つこれらの特性を生かし,漠然とした情報要求を,ユーザが言語によって表現することなく,また,特別なインタフェースを用いることなく検索を行うことを目的とし,クエリの一部として絵文字を受け付ける検索手法の提案を行う. 提案手法では,絵文字を文書検索で用いる方法として,絵文字の持つ多義性を言語に変換するための辞書を用意し,辞書を用いて絵文字を言語に変換し,検索を行うという手法を提案する.

連絡先:山本千尋, NTT サイバーソリューション研究所,
yamamoto.chihiro@lab.ntt.co.jp

2. 関連研究

本研究に関連する技術として,漠然とした情報要求を解決する手段として考えられる技術と,文字列以外のクエリを受け付ける検索技術の 2 つを検討する.

2.1 漠然とした情報要求に対応可能である技術

漠然とした情報要求を解決する手段として考えられる,クエリ拡張技術の研究や,ベクトル空間モデルを用いた研究が行われている.

クエリ拡張技術の一つとしてあげられるのが,シソーラスを用いる手法である. シソーラスの生成方法では,人手でシソーラスを作成する方法や[Voorhees1994],文書データベースに含まれる文書の単語間共起などに基づいてシソーラスを自動生成する方法がある[Qui1993]. その上で,クエリに含まれる単語からシソーラスを引いて関連する単語をクエリに追加することでクエリの拡張を行うというものである. ベクトル空間モデルとは,検索キーワードをベクトル化し,検索文書との距離を計算することで,検索キーワードを含まない文書であっても,類似する文書を検索することが可能な技術である [Salton1983].

本研究は,絵文字を辞書を用いて言語化することから,シソーラスを用いたクエリ拡張技術の一部で,漠然とした情報要求に対応できるように拡張したものであると言える.

2.2 文字列以外のクエリを受け付ける検索技術

文字列以外のクエリを受け付ける検索技術として,画像をクエリとして受け付けるものがある. 例えば,携帯電話のカメラで撮影された画像を用いて文書を検索する技術である[Yeh2005].

この研究では,検索クエリとして入力された画像の持つ属性情報から検索クエリを生成して検索を行なっているため,画像の入力の際に,カメラなど特別なインタフェースを必要とする.

本研究で用いる絵文字は,画像の一部であると言うことができるが,特別なインタフェースを用いることなく,容易に入力することが可能である.

3. 絵文字を入力とする検索手法の提案

クエリの一部として絵文字を受け付ける検索手法の提案を行う. 本研究では,絵文字を文書検索で用いる方法として,絵文字を言語化するための辞書を用意し,検索を行うという手段をとる.

3.1 絵文字辞書の作成

絵文字を言語に変換するために,携帯電話で用いられる絵文字に対し単語を与えるための,絵文字と単語からなる辞書を人手で作成した. 対象とする絵文字は,NTT DoCoMo の携帯電話で利用可能な i-mode 用絵文字 252 種とした.

表1に辞書の例を示す. 表中の“タイトル”は,絵文字を提供している NTT DoCoMo によってつけられた単語である.

今回は,絵文字,もしくは絵文字の“タイトル”から想起される単語を,被験者 3 名によって議論によって決定した. 基準としては,なるべく汎用的になるように,数に際限なく決めてもらう方法をとった. 絵文字から想起される単語を決定する際に,絵文字,もしくは絵文字の“タイトル”から与えられる汎用的なイメージのみを採用し,普段メールなどで用いられる,特定のユーザ間でのみ使用される意味を含む単語は省いた. 例えば,「☀」であれば,汎用的な意味として想起される例として「晴れ」や「太陽」があげられるが,特定のユーザ間でのメールでは,「ごきげん」や,「やったー」といった意味を含め利用されることがある.

表1に辞書の例を示す. 作成された絵文字辞書は,1 つの絵文字に対し,1～13 単語が与えられた. 与えられた単語数の中央値は 4 語で,単語数が少ないものの例としては,「✊」が「バナナ」,「🐧」が「ペンギン」など,絵文字の抽象度が低いものが挙げられ,与えられた単語数が多いものの例としては,「♨」が「よい気分,温泉,温泉宿,風呂,風呂場,浴室,バスルーム,バス,入浴,風呂に入る,風呂屋,銭湯」など,想起される単語が複数品詞によるもので,抽象度の高いものが挙げられた.

表 1　絵文字辞書の例

絵文字	タイトル	単語 1	単語 2	・・・
☀	晴れ	太陽	晴れ	快晴
⚽	サッカー	サッカー	蹴球	フットボール
💗	黒ハート	ハート	心	恋
💠	かわいい	ラブリー	スウィート	花

3.2 絵文字を入力として受け付ける検索手法

次に,3. 1 で作成した絵文字辞書を用いて,絵文字を入力とする検索を行う手法について提案する. 今回は,クエリとして絵文字が入力された際に,絵文字を単語に変換した上でブーリアン検索を行う手法と,絵文字をベクトルに変換しベクトル空間モデルによる検索を行う手法の 2 つを検討する.

(1) 絵文字を用いたブーリアン検索手法

絵文字を用いたブーリアン検索手法について説明する. ブーリアン検索手法は,ブーリアン演算子を用いて検索クエリを生成し,検索を行う手法で,Google や Yahoo 等で広く採用されている. この手法は,クエリとして入力された絵文字を単語に変換することで容易に利用可能であると考えられる.

① 絵文字と文字列による入力から,絵文字部と文字列部を分割する.
② 絵文字部は,絵文字辞書の単語すべてに変換する.
③ 文字列部と,単語に変換された絵文字部をあわせ,ブーリアン演算子を用いたクエリを生成する. ユーザの情報要求は,絵文字辞書から与えられる単語のいずれか,あるいは,複数の組み合わせで表現されると考えられるため,絵文字部についての検索は,ブーリアン検索における OR 検索であるとした. また,文字列部については,絵文字の多義性に制約をつける制約条件としての役割があると考え,文字列部と,絵文字部は,ブーリアン検索におけるAND検索であると考える.
④ 生成されたクエリによって,検索を行う.

図1に,本検索システムの概要を示す.

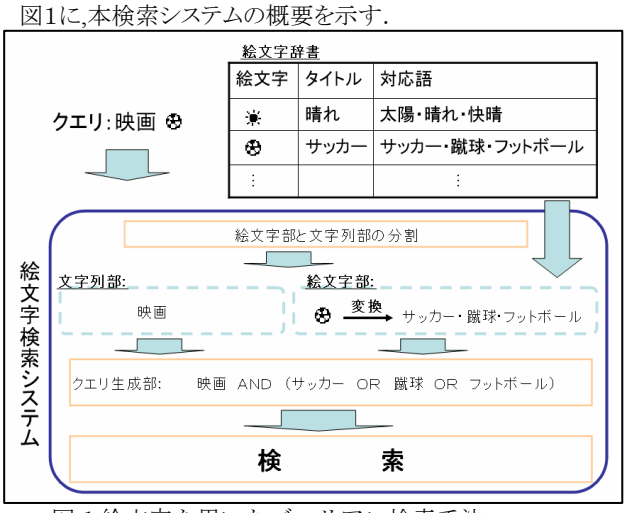

図 1 絵文字を用いたブーリアン検索手法

(2) 絵文字を用いたベクトル空間モデルによる検索手法

絵文字を用いたベクトル空間モデルによる検索手法について説明する. 2.1 で述べたように,ベクトル空間モデルは,漠然とした情報要求に対応する技術であり,漠然とした情報要求を入力する絵文字による検索において有効であると考える.

① 絵文字と文字列による入力から,絵文字部と文字列部を分割する.
② 絵文字部は,絵文字辞書によって言語化する.
③ 文字列部と合わせて,単語それぞれをベクトル化する. それぞれの単語をベクトル化するために,今回は,概念ベース技術[別所 2008]によってあらかじめ与えられる各単語のベクトルを,単語のベクトル化に用いた. 概念ベース技術は,単語と単語意味属性とのコーパス中における共起頻度に基づき,単語のベクトルを生成するものである.
④ それぞれの単語のベクトルの重心を求めることで,クエリベクトルを生成する.
⑤ 生成されたクエリベクトルを用いて検索を行う. 検索文書は,④の概念ベース技術によって単語それぞれをベクトル化し,その重心を求めることで,文書ベクトルを与えておく. 検索では,各文書ベクトルとクエリベクトルの距離が近いものを検索する.

図 2 に,本検索システムの概要を示す.

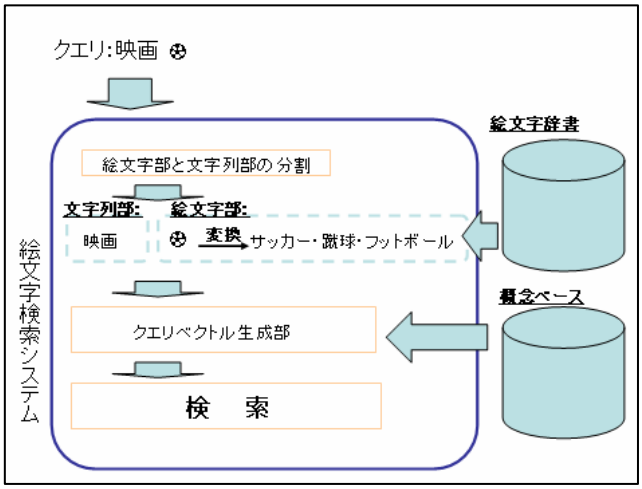

図 2 絵文字を用いたベクトル空間モデルによる検索手法

4. 実験

3. の提案手法に基づいて,絵文字を入力として受け付ける検索システムの実装を行い,既存の単語による検索と比較して,漠然とした情報要求に対応した検索が行えることを確認した.

実験では,提案手法の絵文字による検索結果が,従来の単語による検索結果に比べ,絵文字から想起される範囲で幅広い内容であるかを比較した.

4.1 比較手法

3.2 の(1)における,絵文字を単語に変換し,ブーリアン検索を行う手法と,3.2 の(2)における,絵文字をベクトル化し,ベクトル空間モデルによって検索を行う手法の 2 つを実装した.

比較手法については,絵文字を用いたブーリアン検索手法と,絵文字を用いたベクトル空間モデルによる検索手法,それぞれに対応する手法を実装した.

絵文字を用いたブーリアン検索手法との比較には,単語を入力とし,単語とのマッチングで文書検索を行う検索システムを用いた. 比較に用いる検索クエリとしての単語は,漠然とした検索要求を検索する際に,最も適していると考えられる単語を用いた. 例えば,情報要求が“かわいらしい”という場合であれば,“かわいい”が単語として入力されるのが適切であるとした.

絵文字を用いたベクトル空間モデルによる検索手法との比較には,単語を入力とし,入力された単語をベクトル化して,検索を行う検索手法を用いた. 単語をベクトル化する際に,絵文字を用いたベクトル空間モデルによる検索手法と同様に概念ベース技術[別所 2008]を用いた. ベクトルを用いた検索の際には,提案手法と同様に,検索文書に文書ベクトルを与え,各文書ベクトルとクエリベクトルの距離が近いものを検索した.

4.2 実験データ

評価に用いる対象データは,映画のタイトルと,映画のタイトルを含めた検索クエリで検索された映画に関するブログのスニペットのセット,5966 個を用いた. 表 2 にデータの例を示しした.

実験では,ブログのスニペットを一文書とし,文書部分のみを検索対象として用いた.

表 2　対象データの例

タイトル	文書
星の王子様	・・な気分になったので,“星の王子様”を見ました. 一番好きなシーンは・・・
スパイダーマン	・・久しぶりに,映画館に行って,スパイダーマンをみてきたよ. すっごい・・・

4.3 実験内容

今回は対象データが映画であるため,映画を検索する際の情報要求として考えられる,漠然とした 2 つの情報要求を想定した.

情報要求 1 では,映画の文書に対してかわらしい映画を検索することを目的とし,絵文字「」で検索した結果と,単語「かわいい」で検索した結果を比較した. 絵文字「」は,絵文字辞書によ

表 3　実験結果

	ブーリアン検索手法		ベクトル空間モデルによる検索手法	
	情報要求1「」・かわいい	情報要求2「」・恋愛	情報要求1「」・かわいい	情報要求2「」・恋愛
単語による検索	・プリティ・ウーマン ・プリティ・ブライド ・ウーマン・オン・トップ ・ラブソングができるまで ・スパイダーマン ・めぐり逢えたら ・17 歳のカルテ ・ティファニーで朝食を ・星の王子さま ・ジュマンジ	・幸せになるための 27 のドレス ・ウィンブルドン ・魔法にかけられて ・ラブソングができるまで ・グリーン・ディスティニー ・ラッキー・ユー ・最後の恋のはじめ方 ・オール・ザ・キングス・メン ・デスパレートな妻たち ・キル・ビル	**1.プラダを着た悪魔** **2.プラダを着た悪魔** **3.リトル・ミス・サンシャイン** 4.スパイダーマン 5. ハンコック 6.ジュマンジ 7.スパイダーマン 8.幸せのレシピ **9.つぐない** **10. つぐない**	1.ラブアクチュアリー 2.魔法にかけられて 3.グリーン・ディスティニー 4.ホリデイ 5.タイタニック 6.ホリデイ 7. ホリデイ 8. 魔法にかけられて 9.タイタニック 10.プライドと偏見
絵文字による検索	・マグノリアの花たち **・めぐり逢えたら** ・幸せになるための 27 のドレス ・プリティ・ウーマン ・プリティ・ブライド ・ウーマン・オン・トップ ・ラブソングができるまで ・オール・ザ・キングス・メン ・ホリデイ ・17 歳のカルテ	・シティ・オブ・エンジェル ・ボーンコレクター ・チャーリーズ・エンジェル ・マグノリアの女たち ・プラダを着た悪魔 ・アメリカン・スウィートハート ・噂のアゲメンに恋をした！ ・ベスト・ブレンズ・ウェディング ・ラブソングができるまで ・ラブアクチュアリー	**1.プラダを着た悪魔** **2.プラダを着た悪魔** 3.フック 4.フック 5.星の王子様 **6. リトル・ミス・サンシャイン** 7.サンキュー・スモーキング **8.つぐない** **9.つぐない** 10.ホリデイ	1.ロンリーハート 2.ブレイブハート 3.アメリカン・スウィートハート 4.ラブアクチュアリー 5.ラブ・オブ・ザ・ゲーム 6.ラブ・オブ・ザ・ゲーム 7.クリスティーナの好きなコト 8.ラブソングができるまで 9.幸せになるための 27 のドレス 10.ラブアクチュアリー

って,「かわいい,ラブリー,スイート,チャーミング,花,フラワー」の 6 単語に変換された.

情報要求 2 では,映画の文書に対して恋愛物の映画を検索することを目的とし,絵文字「💔」で検索した結果と,単語「恋愛」で検索した結果を比較した. 絵文字「💔」は絵文字辞書によって,「ハート,ハートマーク,心,愛,恋,ラブ」の 6 単語に変換された.

4.4 結果

実験の結果を表 3 に示した. 検索結果として,文書部分を検索対象として検索されたデータセットの,映画のタイトル部分を示した.

表の右側に,ブーリアン検索手法の,単語による検索と絵文字による検索の検索結果を 2 種類のクエリそれぞれについて示した. 表の左側にベクトル空間モデルによる単語による検索と,絵文字による検索の検索結果を,2 種類のクエリそれぞれについて示した.

ブーリアン検索による結果では,検索結果の例として 10 個の結果を出力した. ベクトルを用いた検索による結果では,検索クエリのベクトルから距離が近かったものから順に 10 件の検索結果を示した.

表 4 にブーリアン検索における,単語による検索と,絵文字による検索の検索結果数を,2 つの情報要求それぞれについて示した. 2 つの情報要求で,ともに,単語による検索に比べ,絵文字による検索の結果件数が多い傾向があることが分かった. 絵文字による検索では,絵文字を単語に変換する際に単語数が増えるため,ブーリアン検索を行う際に,検索結果件数も増加したと考えられる.

表 4　実験結果

	情報要求1「⛲」・かわいい	情報要求2「💔」・恋愛
単語による検索	10 件	15 件
絵文字による検索	21 件	195 件

ベクトルを用いた検索結果では,情報要求 1 では,単語による検索結果と,絵文字による検索結果が, 表中太字の部分で示すように,上位 10 個中 5 件が同じ検索結果だったことから,非常に類似していることが分かった. 違いとしては,単語による検索では,検索対象文書中に,“かわいい”が含まれたものが上位になったが,絵文字による検索では,表中下線部で示す,検索対象文書中に“チャーミング”を含む「フック」や, “花”を含む「星の王子様」など,“かわいい”を連想させる単語が含まれたものが上位に含まれた. 情報要求 2 では,単語による検索では,検索対象文書中に,“恋愛”が含まれたものが上位になったが,絵文字による検索では,検索対象文書中に“恋”を含む「ロンリーハート」や,“ラブ”を含む「アメリカン・スウィートハート」など,“恋愛”を連想させる単語が含まれたものが上位に含まれた.

4.5 考察

実験の結果,ブーリアン検索手法では検索キーワードの拡張により,検索結果がより多岐にわたり,漠然とした情報要求を網羅することができることが分かった. 表中の「⛲」の検索結果である,表中下線部の, “めぐり逢えたら”は,検索対象文書中に“かわいい”という単語は含んでおらず,“花束”というキーワードによって検索されていた.

また,ベクトル空間モデルによる検索手法による検索では,単語による検索と,絵文字による検索において結果が非常に類似する場合があることが分かった. しかし,図 3 に示すように,単語による検索では,ユーザは漠然とした情報要求を持つ場合に,いったん言語化する作業を強いられるが,絵文字による検索では,漠然とした情報要求に対応する絵文字を想起できれば,絵文字をそのままクエリの一部として用いることができるので,より漠然とした情報要求のままクエリとすることができる.

5. まとめ

本研究では,絵文字を検索クエリとして用いることで,漠然とした情報要求を,ユーザが言語によって表現することなく,また,特別なインタフェースを用いることなく検索を行うことを目的とし,クエリの一部として絵文字を受け付ける検索手法の提案を行った. 提案手法では,絵文字を文書検索で用いる方法として,絵文字を言語化するための,絵文字辞書を作成した. さらに, 絵文字を単語に変換した上でブーリアン検索を行う手法と,絵文字をベクトルに変換しベクトル空間モデルによる検索を行う手法のそれぞれを,従来の単語による検索と比較し,絵文字による検索手法の結果が,従来の単語による検索結果に比べ,絵文字から想起される範囲で幅広い内容であるかを比較した.

今後は,検索結果の評価と,ユーザによる主観評価を行う予定である.

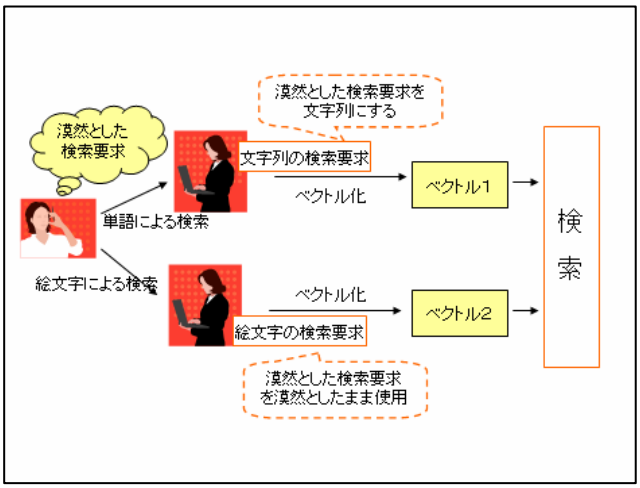

図 3 絵文字による検索と単語による検索の比較

参考文献

[Voorhees1994]Voorhees, E. M.: “Query expansion using lexical-semantic relations”, Proc. Of ACM-SIGIR’94,pp.61-69, 1994.

[Qui1993]Qui, Y., Frei, H.P.,: “Concept Based Query Expansion”, Proc. Of ACM-SIGIR’93,pp.160-169, 1993.

[Salton1983]Salton, G., MacGill, M. J.: “Introduction to Modern Information Retrieval”, MacGraw-Hill,, 1983.

[Yeh2005]Yeh, T., Grauman, K., Tollmar, K., and Darrell, T.: “A picture is worth a thousand keywords: image-based object search on a mobile platform”, CHI ’05, pp.2025- 2028, Portland, ore, USA April 2005

[別所 2008] 別所克人,内山俊郎,内山匡,片岡良治,奥雅博: “単語・意味属性間共起に基づくコーパス概念ベースの生成方式”, 情報処理学会論文誌,Vol. 49 No.12 pp.3997-4006, 情報処理学会 , 2008