タスク主導検索におけるリスク情報のQAコーパスからの発見

北口　善紀†　　大島　裕明††　　田中　克己††

† 京都大学工学部情報学科　〒606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科社会情報専攻　〒606–8501 京都府京都市左京区吉田本町
E-mail: †{kitaguchi,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　本論文では，QA コーパスからリスク情報を発見する手法を提案する．リスク情報とは，将来的に被害に遭う可能性があることを示す情報のことである．品物の購入やサービスの利用といったタスクを行う際に，リスク情報を提示することで，タスクを行うユーザの知識を増やすことが可能となる．リスク情報の発見には，QA コーパスを用いる．QA コーパスとは質問応答サイト（QA サイト）のコンテンツの集合体である．QA コーパスをリスク情報のもつ性質を利用してマイニングすることで，効率よくリスク情報を取得することが可能となる．

キーワード　タスク主導検索, リスク情報, QA コーパス, 意思決定

1. はじめに

近年，情報検索技術の向上により，ユーザが求める情報を取得するのは，以前と比べて遥かに容易になっている．しかし，依然として取得しづらい情報が多々あるのも事実である．取得しづらい情報の一つとしては,「タスク主導検索におけるリスク情報」が挙げられる．まずは,「タスク主導検索におけるリスク情報」がどのような情報であるかを説明する．

タスク主導検索とは,「あるタスクを行う際に必要となる検索」であると本研究では定義する．本研究で対象とするのはタスクに関係する意思決定である．意思決定とは「複数の選択肢の中から，一つ（ないしは複数）の選択肢を選ぶこと」である [8].「洗濯機を購入する」というタスクを例として考えると,「メーカーを決める」,「タイプ（縦型かドラム式か）を決める」といった意思決定がタスクを行う上で必要となる．意思決定が関わる他のタスクとしては「洗濯機を購入する」,「ハワイへ旅行に行く」,「予備校を選ぶ」などが例として挙げられる．意思決定を行う上で有意義な情報は，タスク主導検索においても有意義であると考えられる．意思決定を行う上で有意義な情報の一つとして，リスク情報が挙げられる．

リスク情報とは,「将来的に発生する可能性のある被害に関する情報」である．本研究では行動として意思決定を対象にするので,「意思決定においてある選択肢を選んだ際に，将来的に発生する可能性のある被害に関する情報」と言い換えることができる．意思決定は以下のステップで行われるべきであるとされている [8].

（1） 問題の定義
（2） 評価基準の発見
（3） 基準間の重み付け
（4） 選択肢の生成
（5） 基準に基づいた選択肢の評価
（6） 最適な選択肢の計算
（7） 選択肢の選択

本研究では，評価基準の発見というステップに利用できる手法を提案する．重要な評価基準を発見できなかったために，選択肢の評価を誤り，不適切な選択肢を選択してしまうということが意思決定においてあり得る．評価基準として発見できない事が多い情報としてはリスク情報が挙げられる．通販で物を買う上でどのサイトを利用するか決めるというタスクを例として考える．メリットとなる情報は，通販サイト内でわかりやすく記述されていると考えられるので，簡単に取得することができる．それに対して，デメリットとなる情報は通販サイトでわかりやすく記述されていることは少ないと考えられるため，簡単には取得できない場合がある．本論文でのリスク情報も，デメリットとなる情報に当たる．そのため，リスク情報の収集が十分でないままに選択を終えてしまった場合，あとでリスク情報が指し示す被害に遭い，選択が失敗だったと後悔するというパターンは往々にして考えられる．このような失敗を避けるためにも，リスク情報を発見することは有意義である．

タスク主導検索においては，リスク情報はタスクに関わる意思決定の選択肢とのペアで発見されるのが望ましいと本研究では考えている．リスク情報は，本来タスクに依存するものであるが，意思決定の選択肢に選んだものに関わらないリスク情報は取得できても，意思決定を行う上ではあまり役に立たないと考えられる．ヘッドフォンの購入というタスクを例に挙げる．このとき,「ヘッドフォンを使うと耳が悪くなる」というリスクに関わる情報が得られても，どのヘッドフォンを買えばリスクを避けることができるのかを判断することはできない．それに対して,「密封式のヘッドフォンは耳が悪くなる可能性が高い」という情報が得られたとする．この場合，密封式以外のヘッドフォンを購入すれば，リスクを軽減できることが理解できる．そのため，ヘッドフォンの購入というタスクにおいて,「密封式のヘッドフォンは耳が悪くなる可能性が高い」という情報は有用であると言える．そのような理由から，リスク情報は，それに関わる意思決定の選択肢と共に取得されるべきなのである．

本研究では，タスクを入力として与えたときに，意思決定の選択肢とリスク情報のペアが取得できる手法を提案する．リスク情報の取得には，選択肢に関わるリスク要因を用いる．まず

選択肢とリスク要因のペアを取得し，それらを入力とすることでリスク情報を取得できる手法を用いてリスク情報を取得する．リスク要因とは選択肢の属性のうちで，リスク情報に関わるものである．

リスク情報の例として，表 1 を参照されたい．ここでの「選択肢」とは，タスクを行う上で考えられる意思決定に関する選択肢である．パソコンの購入というタスクを例に挙げる．このタスクでは，最終的に選択するのは一つのパソコンであるが，どのパソコンを買うか決める前に，どのメーカーのパソコンを買うかという意思決定が必要になることも考えられる．そのため，パソコンだけでなくメーカーも，このタスクの選択肢となりうる．

リスク情報の発見は容易ではない．理由の一つとして，リスク情報は多くの観点での情報が必要となることが挙げられる．あるタスクを行う上で，何がリスクとなるかを知っておくことは重要である．また，不測の事態に対応するためには，数多くのリスクについて知っておくことが必要である．そのため，いろいろな観点から，どのようなリスクがあるかを調べる必要があるが，効率的にリスク情報を収集する手法については，あまり考えられていないため，数多くのリスク情報を収集するのには，時間や手間がかかる．そのため，リスク情報を効率的に取得する手法が必要となるのである．本研究ではリスク情報を QA コーパスから取得する方法を提案する．QA コーパスから取得するには，リスク情報のもつ性質を利用する．

2. 関 連 研 究

本研究の目的はタスク主導検索において，関連する意思決定の選択肢を選ぶ上で重要なリスク情報を取得することである．本研究に関連する研究分野としては，属性情報抽出，評判情報抽出が挙げられる．

評判情報を一般的な Web 検索で得られた膨大な量の検索結果から取得することは難しい．この問題に対処した，一般的な文書からの評判情報を抽出する手法に関する研究としては，Huang [1]，Thet [7] らの研究がある．評判情報の研究分野においては，評判情報内に含まれる属性についてのスコア付けを行うことも重要である．スコア付けに関する研究としては，Scaffidi [6]，菊池 [10] らの研究がある．菊池らの研究では，評判情報の内容の要約手法も提案している．レビューを用いた検索手法を提案する研究としては，杉木ら [11] の研究がある．

属性情報抽出の分野でも様々な研究が行われている．Ravi [5] らは，構造化された Web 中の文書からクラス属性を抽出する手法を提案した．Putthividhya [4] らは商品のタイトルのリストから，商品の属性と属性値を抽出する手法を提案した．

本研究は意思決定における選択肢とそれに関係するリスク情報を取得することが目的である．選択肢をどう取得するかという問題は，属性情報抽出の分野に関わりがあり，リスク情報をどう取得するかという問題は，評判情報抽出の分野に関わりがある．しかし，本研究では選択肢とリスク情報の組を取得するということを目的としているため，その点で属性情報抽出，評判情報抽出とは異なる．

3. リスク情報に関わる概念

本章では，本研究に関わる重要な概念について説明する．本研究において重要な概念として，以下の 3 項目が挙げられる．

（1） リスク
（2） リスク情報
（3） リスク要因

以下で，これらの概念一つ一つについての説明を行う．

3.1 リ ス ク

リスク (risk) の定義にはさまざまあるが，一般的には,「ある行動に伴って（あるいは行動しないことによって），危険に遭う可能性や損をする可能性を意味する概念」 と理解されている(注1)．ここで，リスクの定義の中に危険という言葉が含まれているため，危険という言葉の定義が必要になる．危険（きけん，英: Danger，独: Gefahr）とは，未来において，損害や損失が発生する可能性があることである(注2)．本研究では，リスクの定義文中の行動として，意思決定において選択肢を選ぶという行動を対象にする．そのため，本研究では，意思決定においてある選択肢を選ぶと，損害や損失が将来的に発生する可能性がある状態になることをリスクがある行動として定義する．本研究ではこのようにリスクの定義するため，意思決定においてある選択肢を選ぶと必ず損害や損失があるという場合には，その選択肢を選んだ状態はリスクのある状態ではないと考える．

3.2 リスク情報

リスク情報の持つ性質として，

（1） 危険性
（2） 不確実性

があると本研究では考える．本研究でのリスク情報は，上記の二つの性質を必ず持たなければならない．

3.2.1 危 険 性

本論文でのリスク情報は被害が起こりうることを表す情報であるため，危険性があることが分かる情報でなければならない．危険とは，未来において，損害や損失がある可能性があることである．本研究では，その情報を見ただけでどのような損害や損失があるか理解できる情報のみをリスク情報として扱う．パソコンの購入というタスクを例に挙げる．この場合,「メーカー A のサポートが悪い」という情報により,「修理がうまくいかない可能性がある」という危険性があることが理解できる.「修理がうまくいかない」という損害が起きる可能性を示しているため,「サポートが悪い」という情報はリスクを表す情報であると言える．しかし,「サポートが悪い」という情報は単なる属性の情報であり，それ自体がどのような損害があるかを示す情報であるとは言えない．そのため，本研究では「サポートが悪い」という情報はリスク情報には含めない.「サポートが悪い」という情報は，後述するリスク要因に関係する．

3.2.2 不 確 実 性

本研究でのリスク情報は不確実性を持たなければならない．

(注1)：http://ja.wikipedia.org/wiki/リスク (2013 年 1 月 31 日閲覧)
(注2)：http://ja.wikipedia.org/wiki/危険 (2013 年 1 月 31 日閲覧)

表 1　リスク情報の例

タスク	選択肢	リスク情報
ヘッドフォンの購入	密封式	耳が悪くなる可能性が高い
クロスバイクの購入	エンド幅 135mm	交換できるホイールの種類が少ない
パソコンの購入	MacBook Air	メモリの増設ができない
パソコンの購入	メーカー A	修理に手間がかかる
パソコンの購入	メーカー B	故障することが多い
パソコンの購入	海外メーカー	必要最低限のソフトしか付いていないことが多い
パソコンの購入	ネットの激安ショップ	延長保証などがないケースが多い
掃除機の購入	海外版製品	変換アダプタや変圧器が必要となることがある
洗濯機の購入	ドラム式	毛玉がつきやすい

不確実性とは，ある事象が起きることが確実でないことである．不確実性をリスク情報が持たなければならない理由は，本研究において，リスク情報は被害に遭う「可能性」を示す情報と定義しているからである．

不確実性をもつ事象は，ある条件を満たした場合に起こる事象と言える．例として，パソコンの購入というタスクにおいて，MacBook Air という選択肢を選んだ状態を考える．このとき，MacBook Air に関係する情報として,「CD から音楽を読み込む場合には，外付け CD ドライブの購入が必要となる」という情報が考えられる．この情報の,「CD から音楽を読み込む場合」という部分が条件を表していると考えられる．この条件を満たす場合には，外付け CD ドライブの購入を行わなければならないが，条件を満たさない場合には，外付け CD ドライブを購入しなければならないわけではない．そのため,「外付け CD ドライブの購入が必要となる」という情報は，MacBook Air を購入する際に必ずしも必要なことではないため，不確実性のある情報であると言える．

3.3　リスク要因

リスク要因とは，リスク情報に含まれる被害が起こる原因となる性質や属性のことである.「パソコンの購入」というタスクにおいて,「修理に手間がかかる」というリスク情報の損害の原因の一つとして,「サポートが悪い」という事が考えられる．このような原因を表す情報に含まれる属性を，本研究ではリスク要因と呼ぶ.

意思決定において選択肢を決定するための指標として，選択肢が持つ属性が挙げられる．選択肢がもついろいろな属性について，他の選択肢との比較を行い，優れた選択肢を選ぶというプロセスが意思決定においては重要となる．1 章で述べたように，意思決定の失敗の原因の一つとして，評価基準の発見が十分でないということが挙げられる．選択肢がもつ属性も，評価基準に含まれると考えられる．適切な評価基準で選択肢を選ぶためには，選択肢がもつ属性の中でも，悪い属性値をもつことが多い属性を知ることが必要であると考えられる．そのような属性が，本研究でのリスク要因である．

選択肢のある属性の値が悪い場合，その属性により，何らかの損害や損失を受ける可能性がある．この点において，ある属性の値が悪いという情報が起こるかどうかわからない被害，つまり不確実な被害の原因となっていると言うことができる．その点で，属性値が悪いことがリスクにつながる属性はリスク要因と呼べるのである.

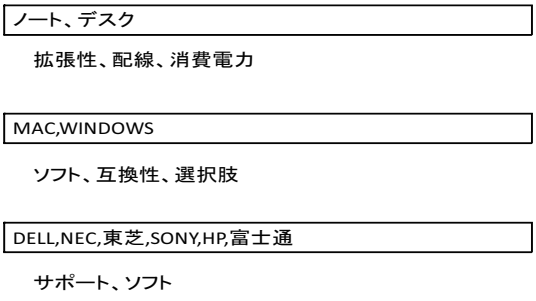

図 1 「パソコン　購入」というクエリでの出力例

本研究ではリスク情報を取得する際にリスク要因を利用する．選択肢とリスク要因を入力とすることで，リスク情報を取得することができる手法を提案し，その手法を用いることで，本研究では選択肢とそれに関わるリスク情報を取得するというアプローチをとった.

4.　リスク情報の取得方法

本章では，入力されたタスクについて，関係する意思決定の選択肢とリスク情報のペアを取得する手法を述べる．入力は，タスクを表す名詞の列であり，タスクについての情報を通常の Web 検索で取得しようときに入力するクエリと同様である.「ヘッドフォンの購入」というタスクについてのリスク情報を得たいときには,「ヘッドフォン　購入」と入力すればよい.

出力は意思決定に関わる選択肢をカテゴリにまとめたもの(以下,「選択肢カテゴリ」と呼ぶ）と，リスク要因語とのペアとする．このペアを入力とすることで，リスク情報を取得する手法も提案する．この手法を用いることで，ユーザが気になった選択肢とリスク要因についてリスク情報を取得できるようにした．図 1 に「パソコン　購入」というタスクを入力とした場合の出力を示す.

本研究では，リスク情報は QA コーパスから取得することとする．QA コーパスから取得する理由は，広範囲の情報を集めることができると考えたからである．実際にリスク情報を取得するという検索意図で 5 件ほどのタスクにおいて Web 検索を行った．その際に，通常の Web 検索で取得したリスク情報が

含まれるページでは同じような内容のものが多かった．また，検索結果に QA が Hit する場合も多く見られた．それに対し，QA で検索した場合には，同内容の質問が投稿されることが少ないため，少ないページビューで多くのリスク情報を取得することができた．そのため，QA コーパスから取得する方が，通常の Web 検索結果から取得するより効率的である，と本研究では考えた．

本研究では，提案手法を実装する上で，QA コーパスとして Yahoo!知恵袋(注3)を用いる．Yahoo!知恵袋とは，Yahoo! JAPAN が運営する，電子掲示板上で利用者同士が知識や知恵を教え合うナレッジコミュニティ，知識検索サービスである(注4)．QA コーパスから質問文書やベストアンサー文書などを取得する際，Yahoo!JAPAN が提供する知恵袋 WebAPI に含まれる，質問検索 API（以下，Yahoo!知恵袋 API と呼ぶ）を実装に用いた(注5)．質問を取得する際のソート項目は，本研究での実装においては指定していない．

以下で，選択肢とそれに関わるリスク情報とのペアの取得方法を述べる．本研究では，下記に示す四つのステップにより，目的の情報を取得する．

（1） 入力されたタスクを表す語の列から選択肢語を取得

（2） 取得された選択肢語を所属する選択肢カテゴリに分類

（3） 選択肢カテゴリについて，関係の強いリスク要因語集合を取得

（4） リスク要因語からリスク情報を取得

以下でそれぞれのステップにおける，手法の詳細について説明する．

4.1 タスクに関わる意思決定の選択肢の取得

タスクに関わる選択肢語を QA コーパスから取得する手法を述べる．選択肢語の取得のために，QA コーパスをクエリを用いて検索し，質問文書集合を取得する．

本研究では選択肢語を多く含む質問文書集合を取得できるクエリがあると考えた．選択肢語を含む質問として，質問者が複数の選択肢を提示し，回答者がおすすめの選択肢を評価基準と合わせて回答してくれることを求める質問がある．このような質問を多く取得できるクエリを追加して検索することで，効率よく選択肢語を取得できると考えた．

本研究では，追加するクエリとして「迷って」という語を用いた．タスクを表す語集合と「迷って」という言葉をつなげたものをクエリとして，QA コーパスを検索することで，N 件の質問文書を取得した．

上記のクエリによる検索により得られた N 件の質問文書を得たのちに，すべての質問文書において，形態素解析を行う．本研究では，手法を実装する上で用いる形態素解析器として MeCab を用いた [2].

形態素解析により，N 件の質問文書から名詞を抽出し，名詞集合 W を生成する．W に含まれる各語 w_i について，選択肢語の重み $\mathrm{s}(w_i)$ を以下の式で与える手法を考案した．

$$s(w_i) = tf(w_i) \cdot \log_2 \frac{N}{df(w_i)} \quad (1)$$

ここで，$tf(w_i)$ は語 w_i が質問文書集合全体で出現した個数であり，$df(w_i)$ は語 w_i が出現する質問文書の数である．

本研究では，選択肢語が含まれやすい文書の中で，選択肢語が含まれる文に共起しやすい語というものが存在すると考えた．「迷って」という語をクエリに追加するという手法では，助詞「か」と動詞「迷う」が含まれる文に選択肢語が含まれることが多いと考えた．そこで，本研究では選択肢語を取得する対象として，質問文書中の「か」または「迷う」という語を含む文を対象にする手法を考案した．

その他に選択肢語を取得するための手法を考案する上で，本研究では，語の特徴が選択肢語に成り得る可能性に関係すると考えた．選択肢語になりやすい語の特徴として，以下の二つを考えた．

- 固有名詞である
- 「型」,「式」などのタイプを表す接尾語を伴う

そこで上記の二つの特徴のいずれか一つでも持つ語は，選択肢語としての重みを多くするという手法を考案した．

4.2 選択肢カテゴリへの分類

選択肢語集合 $C = \{c_1, c_2, ...\}$ について,，選択肢語を所属する意思決定ごとにわけ，選択肢カテゴリ集合 $U = \{A_1, A_2, ..., A_n\}$ への分類を行うためのスコアを求める．A_i は i 番目の選択肢カテゴリを表している．

選択肢語をカテゴリに分類する際に，以下のスコアを用いた．

（1） 共起数

（2） 選択肢語間の最短距離

（3） 共起のバランス

これらのスコアは二つの選択肢語 c_i と c_j について与えられる．本研究では，スコアが高いほど，二つの選択肢語 c_i と c_j は同じカテゴリに属している可能性が高くなるようにスコア付けを行う手法を提案する．

三つの観点でのスコアについて，詳細な説明を以下で行う．

4.2.1 共 起 数

同じ選択肢カテゴリに属する選択肢語は，質問文書の中で並列して記述されることが多くなると考えられる．そこで，並列して記述されている質問文書の数をスコアとすることで，選択肢語が同じカテゴリに属しているかどうかを判別することができると考えられる．

選択肢カテゴリの作成手法でのスコアの算出には，選択肢語をクエリとして得られる質問文書集合を用いる．他の観点でのスコア付けも，同様の質問文書集合を用いる．まず，選択肢語集合に含まれるすべての選択肢語 c_i について，「タスク語 + c_i + 迷って」というクエリで QA コーパスを検索し，N 件の質問文書を含む質問文書集合 Q_i を得る．質問文書集合の取得方法は他の 2 つの観点でも同様である.「パソコン　購入」というタスクにおいて,「ノート」という選択肢については「パソコン　購入　ノート　迷って」というクエリで QA コーパス

(注3)：http://chiebukuro.yahoo.co.jp/

(注4)：http://ja.wikipedia.org/wiki/Yahoo!知恵袋（2013 年 1 月 31 日閲覧）

(注5)：http://developer.yahoo.co.jp/webapi/chiebukuro/chiebukuro/v1/questionsearch.html

を検索し，質問文書集合を得ることになると考えてもらえればよい．

こうして得られた質問文書集合を用いて，選択肢語 c_i と c_j の共起度に関するスコアを以下の式で与える．

$$S_{count}(c_i, c_j) = (df_i(c_j) + df_j(c_i)) \quad (2)$$

ここで，$df_i(w)$ は Q_i のうちで，語 w が含まれる質問文書の個数である．

4.2.2 選択肢語間の最短距離

同じカテゴリに属する選択肢語は同じ文書中に現れる場合に，近い位置に並列することがあると考えられる．これは，複数の選択肢で迷っている質問者が,「A と B で迷っています」などのような質問をし，選択肢語を近い位置に記述する傾向にあると考えたためである．そこで，この手法では，まず質問文書集合 Q_k に属する質問文書のうちで，選択肢語 c_i と c_j を共に含むすべての質問文書を取得し，それぞれの質問文書において，文書内での c_i と c_j の距離を求める．求めた距離のうちで最小となるものを，質問文書集合 Q_k における二つの選択肢語間の最短距離 $D_k(c_i,c_j)$ としてスコアに用いる．

この観点でのスコアは以下の式で求められる．

$$S_{distance}(c_i, c_j) = \max\{\frac{1}{(D_i(c_i, c_j)+1)}, \frac{1}{(D_j(c_i, c_j)+1)}\} \quad (3)$$

上記の式は，質問文書集合 Q_i または Q_j に含まれる文書の中で，c_i と c_j の最短距離の逆数をスコアに用いることを表している．二つの選択肢語が共起している質問文書が存在しない場合には，この観点でのスコアは 0 になる．

4.2.3 共起のバランス

選択肢語 c_i と c_j の共起のバランスを $Balance(i,j)$ とスコア付けする．$Balance(i,j)$ は以下の式で求められる．

$$Balance(i,j) = \min\{\frac{df_i(c_j)}{df_j(c_i)}, \frac{df_j(c_i)}{df_i(c_j)}\} \quad (4)$$

$Balance(i,j)$ というスコアは，上位の選択肢語と下位の選択肢語が同じカテゴリに属することを防ぐために用いる．今回の手法では，df 値の偏りを共起のバランスを表す値として利用している．

$Balance(i,j)$ というスコアを用いる理由を以下で述べる．上位の選択肢語は，下位の選択肢語をクエリとして得られた質問文書集合に多く含まれると考えられる．それに対して，下位の選択肢語は上位の選択肢語をクエリとして得られた質問文書集合に多く含まれるとは限らないと考えられる．そのため，上位と下位で出現数に偏りが見られると予想されるため，$Balance(i,j)$ という値を用いることで，上位と下位の選択肢語が同じカテゴリに属してしまうことを防ぐことができると考えた．そのような理由から，$Balance(i,j)$ というスコアを選択肢カテゴリの分類に関わる手法で用いることとした．

4.3 選択肢カテゴリとの関係が強いリスク要因語の取得

タスクに関係する意思決定の選択肢カテゴリ全体の集合を U とする．U は複数の選択肢カテゴリで構成されているため，$U = \{A_1, A_2, ..., A_n\}$ と表せる．U に含まれる各選択肢カテゴリ $A_i (1 \leqq i \leqq n)$ について，リスク要因語集合 $R_i = \{r_1, r_2, ..., r_m\}$ を求める．

リスク要因語としては，名詞を対象とする．取得先は QA コーパスのベストアンサー文書集合とした．リスク要因語は，タスクに関わる対象の悪い属性を表している語である．そのため，リスク要因語は悪い属性について言及している文書に頻出することが考えられる．QA コーパスを検索する際のクエリに追加することで，悪い属性について言及しているベストアンサー文書が多く得られる語があると本研究では考えた．そのような語の一つとして「デメリット」という語が挙げられる．本研究では,「デメリット」という語をクエリに含め QA コーパス内を検索することによって得られたベストアンサー文書集合に，悪い属性について言及している文書が多く含まれていると考えた．

以下で，選択肢カテゴリ A_k に関するリスク要因語集合の取得方法を述べる．A_k に属する各選択肢語 $a_{ki} (1 \leqq i \leqq |A_k|)$ について，QA コーパスを「タスク語＋ a_{ki} ＋デメリット」というクエリで検索し，上位 N 件に含まれる質問に対するベストアンサー文書を取得する．ここで，$|A_k|$ は選択肢カテゴリ A_k に含まれる選択肢語の個数である．

上記の手法により選択肢カテゴリ A_k からベストアンサー文書集合 BA_k を得る．得られた BA_k からリスク要因語を取得する．選択肢カテゴリ A_k に強く関わりがあるリスク要因語の候補は，BA_k に含まれる名詞である，と本研究では考えた．属性は名詞で表されることがその理由である．BA_k を MeCab を用いて形態素に分割し，名詞である語を抽出する．抽出された名詞である語の集合を W_k とする．

W_k からリスク要因語となる可能性が低い語を除外する．除外するルールとして，以下のパターンに当てはまらない語 w を W_k から除外することとした．

$$\begin{aligned} &w + \text{は} + ((\text{形容詞}) or (\text{形容動詞})) \\ &\qquad\qquad \text{または} \\ &w + \text{が} + ((\text{形容詞}) or (\text{形容動詞})) \end{aligned} \quad (5)$$

このパターンが表しているのは，語 w が形容詞または形容動詞を用いて評価することができる語であるかどうかということである．リスク要因語は悪い属性を表す語であるため，属性値について言及されている可能性がある.「洗浄力」という属性であれば「洗浄力が弱い」,「サポート」という属性であれば「サポートが悪い」，といったように，属性は形容詞で表された属性値をもつことがある．形容詞で属性値が表されているような語を，上記のパターンで取得できると本研究では考えた．W_k からパターンに当てはまらない語を除くことにより得られる語集合を Att_k とする．

上記の手法により得られた Att_k に含まれる語 w_{ki} について，リスク要因語であるかどうかを判定するために以下のスコアを用いる．

$$RiskProbability(w_{ki}) = tf_k(w_{ki}) \cdot \log_2 \frac{N_{all}}{df_{all}(w_{ki})} \quad (6)$$

$tf_k(w)$ は BA_k に含まれる語 w の中で，パターンを満たすも

のの個数, つまり Att_k に含まれる語の個数である．$df_{all}(w)$ はタスクに関係するすべての選択肢カテゴリから取得されたベストアンサー文書集合の中で，パターンを満たす語 w を含むベストアンサー文書の個数であり，以下の式で表すことができる．

$$df_{all}(w) = \sum_{i=1}^{n} \mid \{x \mid w \in W(a) \wedge a \in BA_i\} \mid \quad (7)$$

ここで，n はタスクに関わる選択肢カテゴリの個数を表し，$W(a)$ はベストアンサー文書 a に含まれる語の集合を表している．また，N_{all} はタスクに関わるすべての選択肢カテゴリから得られたベストアンサー集合に含まれるベストアンサー文書の個数であり，n を用いると以下の式で表すことができる．

$$N_{all} = \sum_{i=1}^{n} N \mid A_i \mid \quad (8)$$

Att_k に含まれる全ての語 w について式 (6) によりスコア付けを行ったのちに，スコアが高い順に並べ，上位 m 件に含まれる語をリスク要因語とした．このようにして，選択肢カテゴリ A_k に関係が強いリスク要因語集合 $R_k = \{r_1, r_2, ..., r_m\}$ が得られる．

4.4 リスク要因語からのリスク情報の取得

上記の手法により，選択肢語とリスク要因語がわかれば，リスク情報を取得することができる．リスク情報の取得には，TextRank アルゴリズム [3] を用いる．TextRank アルゴリズムとは，PageRank アルゴリズムをテキストの分野に応用したもので，テキスト間の類似度を用いることにより，文書集合の要約となっている文書を取得するといったことが可能になる．

リスク情報の取得には，ベストアンサー文書を用いる．まず，「選択肢語+リスク要因語」というクエリで QA コーパスを検索し，ベストアンサー文書集合を得る．次に，集合中のベストアンサー文書の中から，リスク要因語を含む文をすべて取得する．そうして得られた文の集合を $S = \{s_1, s_2, ..., s_n\}$ とする．本研究では，文 s_i と文 s_j の類似度を以下の式で与える．

$$Similarity(s_i, s_j) = \frac{\mid w \mid w \in W(s_i) \wedge w \in W(s_j) \mid}{\log_2 \mid W(s_i) \mid + \log_2 \mid W(s_j) \mid} \quad (9)$$

ここで，$W(s_i)$ は文 s_i に含まれる語の集合を表し，$\mid W(s_i) \mid$ は語集合 $W(s_i)$ に含まれる語の個数を表している．文の間の類似度により，TextRank を求め，TextRank の値が最も高い文をリスク情報を含んでいる文として取得する．

5. 実　　験

本章では提案手法を実装し，提案手法の四つのステップのうちリスク情報取得手法を除いた 3 ステップにおいて，提案手法の評価のための実験を行った．実装には C#を用いた．ステップごとに，実験の結果を記述する．

5.1 選択肢の取得

本研究で手法の有用性を示すため，以下の四つの手法を用いて取得した選択肢語集合の上位 20 件での平均適合率を測定した．なお，今回の実験では質問文書の取得数 N については，$N = 100$ とした．N の値は他の 2 ステップにおいても同様に $N = 100$ とした．適合の判断は，選択肢語であると判断できる名詞を適合とした．

（1） タスク語の列のみでの検索により得た質問文書集合から tf-idf 値の高い名詞を取得 (tf-idf 手法)

（2） タスク語の列に「迷って」という語を加えたものをクエリとした際に得られる質問文書集合から tf-idf 値の高い名詞を取得 (+迷って手法)

（3） 2 番目の手法で,「か」または「迷う」という形態素を含まない文に含まれる名詞を除外（+文制限手法）

（4） 3 番目の手法に加え，固有名詞または「型」「式」という接尾語を伴う名詞の重みを増加 (+特徴利用手法)

ここで，平均適合率（AP）とは，検索結果で適合文書が得られた時点での適合率の値の平均値のことである [9]．平均適合率の平均をとったものが MAP であり，本研究では評価の基準として MAP の値を用いる．

MAP を評価基準として用いる理由は，本研究では，条件を満たす語が高いスコアを得られる手法を提案しているためである．このステップにおいては，選択肢語が条件を満たす語に当たる．本研究の手法では，語をスコアが高い順に並べた際に上位に条件を満たす語が集中することが望ましいと本研究では考えている．そのようなランキング手法を評価する基準として情報検索の分野では MAP という評価基準が用いられる．そのため，本研究でも MAP を評価基準に用いた．

表 2　選択肢語取得手法ごとの平均適合率と選択肢語取得数

手法	tf-idf	+迷って	+文制限	+特徴利用
MAP	0.75	0.77	0.79	0.80
AP の最小値	0.19	0.57	0.71	0.74
平均取得数	7.5	11.5	10.8	13.3

結果は表 2 のようになった．表中の AP は平均適合率を表していて，AP の値は，小数第 3 位を四捨五入している．MAP は AP の平均値を表している．手法を加えることで MAP の値が上昇していることがわかる．また，+文制限手法の段階ですべてのタスクにおいて，AP の値が 0.70 を超えている．この結果から，本研究の提案手法は選択肢語をタスクに依存しない一定の精度で取得できる手法であると言える．

+迷って手法と+文制限手法を比べると, AP の最小値では+文制限手法が優れているが，+文制限手法では選択肢語の平均取得数が+迷って手法を下回る結果となった．

5.2 選択肢カテゴリへの分類

以下のタスク語の列と，それに関わる選択肢集合で手法の有用性を調査した．なお，これ以降の評価では,「予備校　選び」というタスクでは選択肢のカテゴリが一つしか求められないため，この入力を評価対象から外すこととする．適合の判断は，同じ選択肢カテゴリに属している選択肢語のペアを適合しているものとした．

本研究では，二つの選択肢語が同じカテゴリに属する度合いを数値化する手法を提案している．そのため，選択肢語集合中のすべての選択肢語のペアを，同じカテゴリに属する度合いで

表 3　選択肢語をカテゴリに分ける手法の平均適合率

手法	共起数	最短距離	共起数+バランス
MAP	0.42	0.55	0.50

表 4　選択肢語をカテゴリに分ける手法の平均適合率

手法	パターン	デメリット	パターン+デメリット
MAP	0.60	0.26	0.74
平均取得数	9.8	3.3	7.3

降順にソートしたものの上位に同じカテゴリに属する選択肢語のペアが来ることが望ましい．そこで，このステップにおいては，以下の三つのスコアを用いてソートした選択肢語のペアのランキングの平均適合率を求めた．

（1） 共起数

（2） 選択肢語間の最短距離

（3） 共起数に共起バランスの値を作用させたもの

結果は表 3 のようになった．共起数のみを用いる手法と共起数に共起バランスを作用させる手法とで MAP 値の改善が見られた．しかし，最短距離を用いる手法の MAP 値を共起バランスを用いた手法が下回る結果となってしまった．

5.3　選択肢カテゴリに関係の強いリスク要因の取得

手法の有用性を示すため，以下の三つの手法を用いてリスク要因語の取得を行い，上位 20 件での平均適合率を測定した．この手法においては，属性を表していると考えられる語を適合する語とした．

（1） タスク語の列をクエリとして得られたベストアンサー文書集合から，パターンを満たす名詞を取得し, tf-idf 値が高い順にソート (パターン手法)

（2） タスク語の列に「デメリット」という語を追加したものをクエリとして得られたベストアンサー文書集合から，tf-idf 値が高い名詞を取得 (デメリット手法)

（3） タスク語の列に「デメリット」という語を追加したものをクエリとして得られたベストアンサー文書集合から，パターンを満たす名詞を取得し，tf-idf 値でソート（パターン+デメリット手法）

評価に用いたタスク語の列とカテゴリ分けされた選択肢語は選択肢をカテゴリ分けする手法と同じものを用いている．

結果は表 4 のようになった．デメリット手法の MAP 値をパターン手法が大きく上回る結果となった．また，デメリット手法, パターン手法とパターン+デメリット手法を比較すると，MAP 値が改善されていることがわかる．

6. 考　　察

本章では，前章で実験した各ステップ毎の結果について考察を加える．

6.1　選択肢の取得

選択肢語取得手法においては，タスク語の列をクエリとして得られた質問文から tf-idf 値の高い名詞を取り出すだけでも 0.70 を超える MAP 値が出ている．しかし，tf-idf 手法では，最小 AP 値が 0.19 となってしまっていて，タスクによって，選択肢語の取得精度にばらつきが見られた．AP の値が小さくなったタスクに共通することは，質問者が選択肢をほとんど知らない状態で質問していることが多いということである．このように選択肢語を含む数が少ない質問文書集合しか得られないタスクの場合には,「迷って」のような選択肢語を含みやすい質問文書を得られる語をクエリに追加するという手法が有効となる．「迷って」という語をクエリに追加することにより，最小 AP 値の改善が見られた．

また，+文制限手法でも最小 AP 値の上昇度は高かった．しかし，+文制限手法では探索範囲を狭めているため，平均取得数が+迷って手法よりも減少してしまった．タスクに関する意思決定においては多くの選択肢を知っておくことが大切だと考えられるので，この手法は選択肢語を多数取得したいという観点からは有効でないと言える．

+特徴利用手法では MAP の値や平均取得数は増加しているが，本研究で提案した特徴を持たないような選択肢語が取得できないという欠点があり，実用には向かないと考えられる．+特徴利用手法と同様のアプローチで精度を上げるためには，選択肢語のもつべき特徴を網羅する必要があると考えられるが，現実的なアプローチではない．

6.2　選択肢カテゴリへの分類

このステップでは，すべての手法であまりいい結果が得られなかった．共起数に加え，共起のバランスを考慮した共起数+バランス手法でも，共起数のみを用いた手法よりは改善されているとはいえ，実用に耐えうる平均適合率であるとは言えない．

最短距離手法が，バランスを考慮に入れた共起数+バランス手法よりも精度が高かった．これは，選択肢語のカテゴリ分けにおいては，共起の数よりも共起している際の語間の距離が重要となることを表している．実際に最短距離によるスコアが高かった選択肢語のペアは，文章中で「A か B」といったように並列して近い距離に記述されているものが多かった．

共起のバランスによるスコア付けにより，上位の選択肢語と下位の選択肢語を別のカテゴリに分けられると本研究では考えていたが，実験ではいい結果が得られなかった．これは上位の語の出現数の多さにより，下位の選択肢語とのスコアが共起のバランスを考慮に入れても高いスコアとなってしまったことが原因である．上位と下位を適切に分けることができるように，スコアの算出方法を見直す必要があると感じた．

6.3　選択肢カテゴリに関係の強いリスク要因の取得

パターンにマッチする名詞を取得することにより，リスク要因語の取得が効率的になることがパターン手法とパターン+デメリット手法から結論づけられる．

パターン+デメリット手法の MAP 値はパターン手法より高くなっているが，リスク要因語の平均取得数はパターン手法と比べると，減少していることが表 4 から読み取れる．これは「デメリット」という語をクエリへの追加により，探索範囲が限定されすぎてしまったことが原因であると考えられる．

質問数が多い「パソコン　購入」というタスクにおいては，パターン+デメリット手法でもリスク要因語の取得数が他のタスクと比べて多くなった．そのため,「デメリット」という語をクエリに加えるという手法が有効になると考えられる．

それに対し，質問数が少ない,「ヘッドフォン　購入」や「ハワイ　旅行」というタスクにおいては，パターン+デメリット手法で取得できたリスク要因語の数が少なく，カテゴリによって AP の値にばらつきが見られた．これはクエリを限定しすぎたことにより，検索により取得できたベストアンサー文書の数が少なくなってしまったことが原因であると考えられる．最もベストアンサー文書の取得数が少なかった「ヘッドフォン　ゼンハイザー　デメリット」というクエリでは，検索結果が 1 件も得られなかった．

この問題は，Yahoo!知恵袋 API で得られる検索結果数が少ないことも原因の一つであると考えられる．Yahoo!知恵袋 API で得られる検索結果の数は，実際に Yahoo!知恵袋で検索することによって得られる検索結果の数よりも少なくなる傾向にある．そのため，Yahoo!知恵袋から直接ベストアンサー文書を取得すれば，精度を改善することができると考えられる．

7. 結　　論

本研究では，タスク主導検索におけるリスク情報のもつ性質について定義し，リスク情報を取得する上で有用なリスク要因を定義した．そして，タスクに関わる意思決定の選択肢カテゴリとリスク要因の組を取得する手法と，選択肢とリスク要因からリスク情報を取得する手法を提案した．それらの手法において，情報は QA コーパスから取得することとした．

本研究で解決していない問題として，関係する属性が不明瞭なリスク情報が取得できていないということが挙げられる．「ヘッドフォンの購入」というタスクを例に挙げる．このタスクにおいて,「密閉式のヘッドフォンは耳に悪い」という情報は,「耳に悪い」というリスク情報を含んでいるため，このような情報も取得できることが望ましい．しかし，本研究の手法では，このような情報はリスク要因語が不明瞭であるため，取得できない．このリスク情報の原因の一つとして，密閉式のヘッドフォンの「遮音性が高い」という特性があり，この文の中には,「遮音性」というリスク要因が含まれている．しかし，遮音性という属性はデメリットとして記述されることが少ないと考えられるため，本研究の手法ではリスク要因として取得することが難しいと考えられる．

また，関係する属性ははっきりしていても，属性以外の条件で記述されているリスク情報を取得することも本研究ではできていない．このようなリスク情報の例としては,「パソコン　購入」というタスクにおいて,「MacBook Air」という選択肢に関係する,「CD から音楽を読み込む場合には，外付け CD ドライブの購入が必要となる」という情報が属性以外の条件で起こる被害に関するリスク情報であると言える．この情報での被害が起こる原因は CD から音楽を読み込むという動作であり，属性が悪い値を持っていることに関係するリスク情報ではない．しかし，この情報もリスク情報として取得できることが望ましい．このリスク情報はリスク要因として「CD ドライブがない」ということが挙げられるため，リスク要因語として取得できる条件を満たす記述があれば取得することはできる．しかし，そのような記述がない場合にも取得できるべきであるため，リスク要因を用いなくともリスク情報を取得できるような手法を提案することも今後の課題である．

謝　　辞

本研究の一部は，文部科学省科学研究費補助金（課題番号 24240013，24680008）によるものです．ここに記して謝意を表します．

文　　献

[1] Shen Huang, Dan Shen, Wei Feng, Catherine Baudin, and Yongzheng Zhang. Improving product review search experiences on general search engines. In *Proceedings of the 11th International Conference on Electronic Commerce*, ICEC '09, pp. 107–116, New York, NY, USA, 2009. ACM.

[2] T. KUDO. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*, 2005.

[3] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004.

[4] Duangmanee (Pew) Putthividhya and Junling Hu. Bootstrapped named entity recognition for product attribute extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 1557–1567, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[5] Sujith Ravi and Marius Paşca. Using structured text for large-scale attribute extraction. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pp. 1183–1192, New York, NY, USA, 2008. ACM.

[6] Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng, and Chun Jin. Red opal: product-feature scoring from reviews. In *Proceedings of the 8th ACM conference on Electronic commerce*, EC '07, pp. 182–191, New York, NY, USA, 2007. ACM.

[7] Tun Thura Thet, Jin-Cheon Na, and Christopher S. G. Khoo. Filtering product reviews from web search results. In *Proceedings of the 2007 ACM symposium on Document engineering*, DocEng '07, pp. 196–198, New York, NY, USA, 2007. ACM.

[8] 印南一路. すぐれた意思決定. 中央公論社, 2002.

[9] 和明岸田. 検索実験における評価指標としての mean average precision の性質. 情報処理学会研究報告. 情報学基礎研究会報告, Vol. 2001, No. 74, pp. 97–104, jul 2001.

[10] 悠太菊池, 大也高村, 学奥村. 属性-評価ペアを単位とした評判情報の要約. Technical Report 1, 東京工業大学, 東京工業大学, 東京工業大学, may 2012.

[11] 健二杉木, 茂樹松原. 消費者の意見に基づく商品検索. 情報処理学会論文誌, Vol. 49, No. 7, pp. 2598–2603, jul 2008.