DEIM Forum 2014 E7-5

リスクとその対策情報の分類と検索

北口　善紀† 　大島　裕明† 　田中　克己†

† 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, リスクをその対策情報の傾向を基に分類し取得する手法を提案する. 本研究の手法でヘッドフォンを利用する際のリスクについて検索すると, 難聴のリスクは事前に対処すべきであり, 故障のリスクは事後に対処するものである, といった情報を得る事ができる. リスクの影響度, 発生頻度によって取るべき対策パターンが異なるという知見を利用し, 対策情報の傾向による分類でリスクを推定することが本研究の狙いである. 本研究では社会的な事故であるトラブルに関するリスクを取り扱う. Web 上に記述されるトラブルやその対策についての情報を基に, 検索クエリで示される状況に即した分類済みのリスクを提示する.

キーワード　リスク, 言語パターン, トラブル, データマイニング

1. はじめに

情報検索技術の向上から私たちは検索意図に合った情報を得られることが多くなったが, 依然として意図通りの結果を得られない場合も存在する. 意図通りの結果を得られない場合の一例として, リスクについての情報を取得したい場合が考えられる. ヘッドフォンを利用する際のリスクについて知りたい場合を例として取り上げる. この場合, 知りたい事柄としては以下の二点が挙げられる.

（1） どんなリスクがあるのか

（2） それぞれのリスクはどの程度の大きさなのか

ヘッドフォンの例では, まず断線するリスクや難聴になるリスク, 音漏れするリスクなどがあることを知ることが重要である. また, 得られたリスクのうち, どれを重要視すべきか知るためにもリスクの大きさについての情報が得られることが望ましい. 本研究では, リスクの大きさを推定するパラメータとして以下の 2 点を考慮する.

- 発生頻度
- 影響度

ヘッドフォンの例について言うと, 難聴になるリスク, 断線するリスクは被害の発生頻度は小さいが, 影響度は大きい. それに対して, 音漏れするリスクは被害の発生頻度は大きいが, 影響度は小さい. 上記のように, それぞれのリスクについて発生頻度, 影響度は異なる.

発生頻度, 影響度をリスクの主要なパラメータとする理由はリスクが「ある状態からダメージを受けた状態へ遷移する可能性がある」場合に定義できるパラメータであるからである. ヘッドフォン購入直後の状態だと, 断線する可能性, 難聴になる可能性, 音漏れする可能性のいずれも存在するため, 断線するリスク, 難聴になるリスク, 音漏れするリスクのすべてについて考慮しておく必要がある. ここでヘッドフォンが断線してしまった場合, ヘッドフォンが断線した状態にユーザの状態は遷移することとなる. この状態ではもうこれ以上ヘッドフォンが断線する可能性がないため, 断線のリスクについては考慮する必要がなくなる. このようにリスクは現在の状態と密接な関わりがある. リスクが状態遷移に付随する情報であると考えたときに, 必要となる情報は以下の二つである.

（1） ダメージを受けた状態へどの程度の確率で遷移するのか (リスクの発生頻度)

（2） その状態に遷移した際のダメージの大きさはどの程度か (リスクの影響度)

以上の理由から, 本研究ではリスクの発生頻度, 影響度について考慮する必要があると考え, それらのパラメータを推定する手法を提案することとした.

現状の検索システムでは解決できない問題として以下の二つが挙げられる.

（1） リスクについての記述があるページを判別できない

（2） リスクの発生頻度, 影響度などによるランキングを作成できない

まず一番目の問題について述べる. ヘッドフォンとして取り上げ, 断線のリスクをすでに知っているユーザがそのリスクにどう対処すべきか知りたいという検索意図で検索を行う事を考える. この場合, 「ヘッドフォン 断線」というクエリで検索することが自然であると考えられる. このクエリで検索した結果として「断線を防ぐためのヘッドフォンの選び方」, 「ヘッドフォンが断線した際の対処法」というタイトルの二つの Web ページが得られたとする. 前者のページは断線する前に読むべきページであり, ダメージを受けた状態に遷移する前に必要な情報について記述しているので, 「ある状態からダメージを受けた状態へ遷移する可能性がある」パラメータであるリスクについての情報を記述していると言える. それに対して, 後者のページではすでに断線したユーザが必要な情報を記述しており, ヘッドフォンが断線した時点ではすでに断線のリスクは消失してしまっているため, リスクについての情報を記述しているとは言えない. これらの二種類のページは「ヘッドフォン 断線」というクエリに合致しているため両方とも検索結果に現れることが予想される. リスクへの対処ついて知りたい場合には必要となるのは前者のページであるが, 前者と後者を分けて取得す

ることは現状の検索では難しいと考えられる.

次に二番目の問題について述べる. ある状態にユーザがいるときに遷移する可能性のあるダメージを受けた状態の数は無数に存在するため, リスクの種類も膨大なものとなる. それらすべてのリスクの情報について理解するのは現実的ではないため, リスクの大きさ等で情報のスコア付けを行い, スコアにより分類された状態で情報を提示する必要があると考えられる. 本研究ではスコアとしてリスクの主要要素である発生頻度, 影響度を利用する. これらのパラメータはリスクを知る上で重要であるが, 容易に推定できるものではないため現状の検索システムでリスクの大きさを推定して情報を提示することは難しいと考えられる. この問題を解決するため本研究では発生頻度, 影響度というパラメータを推定する手法を提案する.

本研究ではリスクの取得･推定にリスクの対策についての記述を利用する. Wikipedia のリスクマネジメントという項目 (注1) では, 「リスク対応の種類には, リスクの回避, 低減, 共有, 保有などがある.」と記述されている. 回避, 低減, 共有, 保有などの対策法はリスクの発生頻度, 影響度に密接な関わりがあると考えられる. そこで本研究では Web 上に記載されている対策の記述を利用し, リスクの取得・推定を行う手法を提案する.

2. 関 連 研 究

本研究は入力されたクエリで指定されたドメインについて, 対策などリスクに関する記述があるページを抽出し, 得られたページ集合からダメージを受けた状態を示す表現を取り出し, 影響度・発生頻度でスコア付けし提示する手法を提案する. 本研究に関連のある研究としてはまず, Walker [3] らの研究がある. Walker らはソフトウェア開発の分野について, リスクマネジメントが困難となっている今日において, テクニカルリスクを客観的に評価し議論できる 3 つの仕組みを提案した. この研究ではイベントについてのリスクを発生頻度とコストで定義していて, その点に関して本研究でのリスクの発生頻度・影響度推定と通づるところがあると考えている. 本研究との相違点は対象としているリスクであり, この論文ではソフトウェア開発におけるリスクを対象にしているのに対して, 本研究では情報検索を利用することの多いドメインにおけるリスクを対象としていることが相違点として挙げられる.

リスクを対象としない関連研究としては, 観点を絞った検索に関する研究が挙げられる. 経験を検索する研究として, 乾 [1] らの研究がある. 乾らはブログなどでユーザが生成したコンテンツのうち, 自分の経験した内容を記述している文書をデータマイニングにより収集する手法を提案している. 本研究でのリスクに関する言及を行っている文書を抽出する際にも, この研究での抽出手法の考え方が利用できると考えている.

3. リスクとその対策

本章では, 本研究を理解する上で重要なリスク及びその対策についての説明を行う.

3.1 リ ス ク

リスクとは, 「ある状態からダメージを受けた状態へ遷移する可能性がある」場合に定義できるパラメータである. 本研究ではリスクの大きさを以下の二つの要素で定義することとする.

（1） 影響度

（2） 発生頻度

リスクマネジメントの分野ではリスクの大きさの影響度と発生頻度の積による定義がしばしば見受けられ, WIkipedia のリスクマネジメトの項目 (注2) でも「リスクマネジメントとは, リスクを特定することから始まり, 特定したリスクを分析して, 発生頻度（発生確率, possibility）と影響度（ひどさ, severity）の観点から評価した後, 発生頻度と影響度の積として求まるリスクレベルに応じて対策を講じる一連のプロセスをいう.」と記述されている. 本研究におけるリスクの大きさの定義にもこの方法が利用できるが, 積での定義の場合には影響度が大きいが発生頻度は小さいリスクと影響度は小さいが発生頻度が大きいリスクが同程度のリスクと評価されることがある. 発生頻度が大きいリスクが求められるか, 影響度が大きいリスクが求められるかはドメインやユーザの嗜好によって異なるため, 影響度・発生頻度をまとめて 1 つのパラメータとして扱うと対応できる要求が限られてしまう. そこで本研究では各々のリスクについて発生頻度・影響度という 2 つのパラメータを別々に推定する手法を提案する.

3.2 リスクの対策

リスクマネジメントの分野では, リスクの対策を以下の 4 パターンに分類している.

（1） リスク回避

リスクのある状態から離脱する

例:難聴にならないようにヘッドフォンを使わないようにする

（2） リスク低減

リスクの影響や発生頻度を少なくする

例:海外旅行時に盗難に備えて財布をズボンのすそに入れておく

（3） リスク共有

リスクによる被害を他の人と共有し分散させる

例:故障に備えてパソコンの長期保証に加入しておく

（4） リスク保有

リスクを黙認し, リスクを減らすためのアプローチを行うことなくリスクを放置する.

例:たまに糸くずがつくことがある洗濯機を使い続ける

リスクへの対策はリスクの影響度・発生頻度によって異なる. リスクの影響度・発生頻度とその対策法の関係を図 1 で示す. 影響度が大きく, 発生頻度も大きいリスクについてはリスク回避が選ばれる. 影響度は小さいが発生頻度が大きいリスクについてはリスク低減が選ばれる. 影響度が大きいが発生頻度は小さいリスクについてはリスク共有が選ばれる. 影響度が小さく, 発生頻度が小さいリスクについてはリスク保有が選ばれる.

上記で示したようにリスクの対策はリスクの発生頻度・影響

(注1)：http://ja.wikipedia.org/wiki/リスクマネジメント

(注2)：http://ja.wikipedia.org/wiki/リスクマネジメント

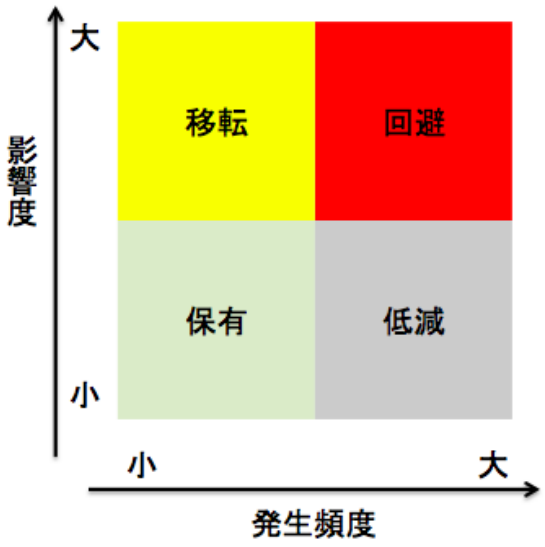

図 1　リスクの影響度・発生頻度とその対策パターン

度というパラメータに密接に関わる．ここで注意しなければならないのが, 影響度・発生頻度を考慮して対策法は選ばれるが, ある対策法が選ばれているリスクだからといって, 影響度・発生頻度が決まるというわけではないということである．ヘッドフォン利用時における難聴になるリスクという影響度が大きいが発生頻度が小さいリスクに対して, リスク回避のためにヘッドフォンを使うときはできるだけ音量を小さくする人もいれば, リスクを保有し大音量で使用しつづける人もいる可能性がある．このようにリスクの大きさが同じでも人によって対策が異なることがある．本研究で対策を基にリスクの抽出・評価を行う際にはこの点を考慮すべきであると考えている．

4.　リスクの抽出・分類手法

本章ではリスクを抽出し, 発生頻度・影響度という観点で評価する手法について述べる．

4.1　リスクの抽出

本研究ではリスクの抽出をリスクを目的語にとる動詞を利用して行う．動詞にターゲットを絞るのは, リスクへの対策が記述してあるページが同じ動詞を含むことが多いという仮説に基づく．例としてヘッドフォンの難聴というリスクへの対策が記述してあるページの特徴に言及する．これらのページでは「難聴の予防法」,「難聴を防ぐには」等の表現が多く用いられている．これらの表現の特徴として「予防」,「防ぐ」等の動詞が「難聴」というリスクを目的語にとっていることが挙げられる．このような表現は決して少ないものではなく, Google を利用して”難聴を予防”というクエリで検索した場合には約 64800 件の結果が得られ, ”難聴を防ぐ”というクエリで検索した場合には約 137000 件の結果が得られたことからも頻出する表現であることが実証できる．本研究では動詞を利用した言語パターンによりリスクを抽出することができるのではないかという仮説に基づき, 以下の言語パターンを利用してリスクを抽出する手法を提案する．

(リスク名) + を + (動詞)　　(1)

ヘッドフォンの例だと言語パターンを満たす表現として,「難聴を予防」,「断線を防ぐ」,「ノイズを解消」などの表現が挙げられる．

実際の動詞を利用したリスクの抽出方法はクエリに動詞を付加したクエリ拡張による検索で行う．ヘッドフォンの例だと「ヘッドフォン　予防」,「ヘッドフォン　防ぐ」等のクエリで検索し, 言語パターンを満たすリスクを抽出する．本研究ではリスクの獲得に有用な動詞を発見するための手法としてブートストラップ法を用いる．統計学におけるブートストラップ法とは, 様々な目的に用いられる統計的推論の手法であり, 再標本化法に分類されるもののひとつである．(注3) 具体的には, まずシードとして利用する動詞集合 $V(0)$ を与え, それを利用して検索を行い, 言語パターンを満たすリスクを抽出し, リスク集合 $R(1)$ を作成する．その次に, $R(1)$ を利用して検索を行い, 言語パターンを満たす動詞を抽出し, 動詞集合 $V(1)$ を作成する．これを何度も繰り返すことで, リスク集合 R とリスク抽出に有用な動詞集合 V を作成する．そうして抽出されたリスク集合 R に含まれる各々のリスク r について, V を用いて分類を行う．以下で, 動詞集合を利用したリスク集合の取得手法, リスク集合を利用した動詞集合の取得手法の詳細を述べる．

4.1.1　動詞集合を利用したリスク集合の取得

動詞集合 $V(t)$ を用いてリスク集合 $R(t+1)$ を求める手法を述べる．ユーザの入力したクエリ q について, $V(t)$ に含まれる動詞 $v \in V(t)$ でクエリ拡張して検索し, 検索結果を 100 件取得する．検索結果の取得には検索 API を利用し, 本研究では Bing の検索 API(注4) を利用している．

各動詞 v について言語パターンによるリスク抽出を行うため, 検索 API で得られたページのタイトルとスニペットを文単位に分割したものをマージした文集合から動詞 v を含む文を抽出する．そうして得られた文集合に含まれるすべての文について言語パターンを満たす部分を抽出する．ここで言語パターンを満たすかどうかを識別するために各文について構文解析を行う．構文解析器には Cabocha(注5) を用いる．ここで形態素解析ではなく構文解析を行うのは, 2 つ以上の形態素が連続して表されるリスクを正しく抽出できるようにするためである．ヘッドフォンに関する音漏れというリスクを抽出する場合だと, 形態素解析を行うと音と漏れが分割されてしまうが, 構文解析では分割されず正しく音漏れを抽出することができる．なお, Cabocha では構文解析のために用いる形態素解析器として MeCab [2] が用いられている．

構文解析により言語パターンを満たすかどうかを識別し, 満たせば動詞 v の前のチャンクに含まれるリスク r について, クエリ q と動詞 v を利用した検索結果におけるリスク r の出現率 $Freq(r, \{q, v\})$ を算出する．q を v により拡張し検索して得られた 100 件の検索結果から r が 35 件得られた場合,

(注3)：http://ja.wikipedia.org/wiki/ブートストラップ法
(注4)：http://datamarket.azure.com/dataset/bing/search
(注5)：https://code.google.com/p/cabocha/

$Freq(r, \{q, v\}) = 0.35$ となる.

抽出されたリスク r が次のステップでの動詞集合の取得に有用かどうかを判断するためのスコア $ER(r, t+1)$ は以下の式で定義される.

$$ER(r, t+1) = \sum_{v \in V(t)} Freq(r, \{q, v\}) \cdot ER(v, t) \cdot EL(v) \quad (2)$$

ここで新たに登場する $EL(v)$ というスコアは, 前回のリスク集合 $R(t)$ と動詞 v により抽出されたリスク集合を比較し, 動詞 v がリスク抽出に有用であるかを測る値であり, 以下の式で表される.

$$EL(v) = \sum_{r \in R(t)} Freq(r, \{q, v\}) \cdot ER(r, t) \quad (3)$$

このスコアを ER(r) を算出する際に用いる理由は, リスク集合を取得するのに有用な動詞から取得されたリスク r が, リスク抽出に有用な動詞の抽出に必ずしも有用であるわけではないという理由に基づく. 例として $q =$ "ロードバイク", $v =$ "防ぐ" から $r =$ "風" が得られた場合を考える. 「防ぐ」という動詞はどのようなクエリに対しても高精度でリスクを抽出できる動詞であることが経験的にわかっているが, この動詞で抽出された「風」というリスクは「感じる」などのリスク抽出に適さない動詞の目的語となることが多いため, 「風」のスコアを上げてしまうと適切な動詞集合を取得することができなくなってしまう. そのため, EL("風") というスコアを求めることで適切な動詞集合が得られやすくする必要があると考え, このスコアを $ER(r, t+1)$ を求める際に利用することとした.

こうして $V(t)$ により抽出されたリスクすべてについて $ER(r, t+1)$ を求めたのち, 次の抽出に利用するリスク集合を $ER(r, t+1)$ のスコアが上位のものに限定する. ここで $ER(r, t)$ を用いてフィルタリングをする理由は, ブートストラップ法で限定なしで反復を続けてしまうと, 集合の大きさが際限なく大きくなってしまうためである. 抽出されたすべてのリスクを用いて次の動詞集合の抽出を行う場合, 計算コストが非常に膨大になってしまうという問題が発生する. それを防ぐため, 本研究では次のステップで利用するリスクを $ER(r, t+1)$ のスコア上位 10 件のものとしている. こうして得られたリスク集合 $R(t+1)$ を用いて, 次は動詞集合 $V(t+1)$ を取得する.

4.1.2 リスク集合を利用した動詞集合の取得

前節で得られたリスク集合 $R(t)$ を利用して動詞集合 $V(t)$ を取得する方法を説明する. 基本的には動詞集合からリスク集合を求める手法と同様である.

まずクエリ q を $r \in R(t)$ で拡張したもので検索を行い, それぞれ 100 件ずつ検索結果を検索 API を利用して取得する. そして各リスク r について得られた 100 件の検索結果のタイトル及びスニペットから, r を含む文を抽出する. 得られた文集合について, 言語パターンを満たす表現があれば, 最初に出現する動詞を抽出し, その頻度 $Freq(v, \{q, r\})$ を算出する. 動詞の取得は形態素解析を利用し, 本研究では形態素解析器として MeCab [2] を利用している. 形態素が動詞であるかどうかの判断は,

（1） 品詞が動詞
（2） 品詞が名詞かつサ変接続

のいずれかの条件を満たすものを動詞として判断することとした.

リスク集合 $R(t)$ に含まれるすべてのリスク r について, 言語パターンを満たす動詞を抽出したのち, 動詞 v のリスク抽出に有用かどうかを表すスコア $ER(v, t)$ を以下の式で算出する.

$$ER(v, t) = \sum_{r \in R(t)} Freq(v, \{q, r\}) \cdot ER(r, t) \cdot EL(r) \quad (4)$$

リスク集合 $R(t)$ から動詞集合 $V(t)$ を求める際にも, 前の動詞集合 $V(t-1)$ を利用したリスク r の動詞抽出への有用性を表すスコア $EL(r)$ を用いる. このスコアは以下の式で算出される.

$$EL(r) = \sum_{v \in V(t-1)} Freq(v, \{q, r\}) \cdot ER(v, t-1) \quad (5)$$

抽出されたすべての動詞 v について, $ER(v, t)$ を求めたのち, $ER(v, t)$ のスコア上位 10 件の動詞を取得し, それを動詞集合 $V(t)$ として取得する.

4.2 リスクの評価

リスク集合 R が抽出されたのち, R に含まれるすべてのリスク r について, r の抽出に最も有用だった動詞 v を動詞集合 V から求める. 選出される動詞 v は以下のスコアが最大となる動詞である.

$$ScoreForSelection(r, v) = Freq(r, \{q, v\}) \cdot ER(v) \quad (6)$$

リスク集合 R に含まれる各リスク r について抽出に有用な動詞 v を動詞集合 V から求めたのち, 得られた動詞 v を用いてリスクを分類する.

分類には, 3 章で述べたリスクの対策が発生頻度・影響度別で 4 パターンに分類されるという知見を用いる. 4 パターンの対策の種類は以下のようになっている.

（1） リスク回避 (発生頻度大・影響度大)
（2） リスク低減 (発生頻度大・影響度小)
（3） リスク共有 (発生頻度小・影響度大)
（4） リスク保有 (発生頻度小・影響度小)

動詞 v が 4 種類の対策パターンのどれに分類されるかは, クエリ q と動詞 v から得られるリスク r の出現率 $Freq(r, \{q, v\})$ による. まず 4 種類の対策パターンを表す代表的な動詞を決めておく. 本研究では下記のように定めた.

（1） リスク回避 ー 「防止」
（2） リスク低減 ー 「軽減」
（3） リスク共有 ー 「補償」
（4） リスク保有 ー 「解消」

リスク r とその抽出に用いられた動詞 v から得られた際に, v が防止, 軽減, 補償, 解消のうちどの動詞との近似度が最も高いかを計算し, 最も類似度が高い動詞の対策パターンにリスク r を分類する. 2 つの動詞 v_1, v_2 の近似度は以下の式で算出される.

	P@10(リスク)	有用動詞の個数	P@10(動詞)
S_1	**0.82**	**1.9**	**0.65**
S_2	0.76	1.5	0.63

表 1　シード動詞集合 S_1, S_2 による抽出結果

$$Sim(v_1, v_2) = \sum Freq(r, \{q, v_1\}) \cdot Freq(r, \{q, v_2\}) \quad (7)$$

この式を用いて, 上記の 4 つの動詞と動詞 v の近似度を計算し, 最も高い近似度となる対策パターンに動詞 v を分類する. こうして分類された動詞 v により抽出されたリスク r については, 動詞 v の対策パターンが適用されることになる. 発生頻度・影響度ごとに対策パターンは異なるため, リスク r が分類された対策パターンが何であるかわかれば, 発生頻度・影響度を評価する事ができる.

5. 実験及び考察

前章で示したリスクの抽出手法について, 10 個のクエリと 2 種類のシードを利用した手法についてリスク, 動詞の抽出実験を行い, リスク集合 R の上位 10 件に含まれるリスクの適合率 P@10(リスク), 上位 10 件のリスクの抽出に有用だった動詞の個数, 動詞集合 V の上位 10 件に含まれるリスクの適合率 P@10(動詞), の 10 個のクエリにおける平均値を求めた. 実験に利用したクエリ及びシードの動詞集合は以下のようになっている.

- クエリ

ヘッドフォン, パソコン, スマートフォン, ロードバイク, ギター, 掃除, 洗濯, 登山, 観光, 海外旅行

- シード

$S_1 =$ { 防止 }, $S_2 =$ { 防止, 軽減, 解消, 補償 }

2 種類のシード動詞集合による抽出を行った理由は, シードの多様性を向上させることにより多様なリスクを抽出できると考えたためである. 実験結果は表 1 のようになった.

すべてのスコアにおいて, シード動詞集合の大きさを大きくした方がスコアが悪くなるという結果となってしまった.

このような結果となった理由として, 1 つの動詞のスコアが高くなりすぎることが多かったことが考えられる. クエリとシード動詞集合ごとに抽出できたリスク集合は表 2, 動詞集合は表 3 のようになっている.

実験前はシード動詞集合に多様性をもたせることにより, 抽出精度が低い「登山」, 「観光」等のクエリに対しても高精度でリスク抽出に適した動詞を取得できるようになり, かつ多様な動詞が抽出できるようになることを期待していた. しかし実際には, リスク抽出に有用な動詞の種類も少なくなり, 適合率も下回る結果となってしまった. 適合率の低下の原因は $ER(v)$ が最も高い v が複数の適切な動詞を抽出するのに適していないものだったことが一因であると考えられる.

また, リスクや動詞について表記揺れを別々に扱っていた事も多様性のスコアを下げる要因になったと考えられる. 表 3 の「洗濯」というクエリにおいてはシード動詞集合に関わらず「落

クエリ	シード	抽出されたリスクと有用動詞
ヘッドフォン	S_1	音漏れ:防ぐ, まり:防ぐ, 振動:防ぐ, 絡み:防ぐ, 絡まり:防ぐ, 侵入:防ぐ, 相互影響:防ぐ, ヘッドフォン難聴:防ぐ, 破損:防ぐ, ノイズ:軽減
	S_2	騒音:低減, ノイズ:低減, 音エネルギー:低減, 雑音:低減, 音漏れ:低減, 音:軽減, 過大突入電流:低減, 負担:低減, 圧迫感:低減, 周辺ノイズ:低減
パソコン	S_1	熱暴走:防止, 情報漏えい:防止, 盗難:防止, 電磁波:防止, 不具合:防止, 情報窃取:防止, 接続:防止, 不正アクセス:防止, 不正使用:防止, 変更:防ぐ
	S_2	疲れ:軽減, 負担:軽減, ブルーライト:軽減, バックアップ:取る, 負荷:かける, 音:とる, 疲れ目:軽減, 疲労:軽減, PC 作業疲れ:軽減, 眼精疲労:軽減
スマートフォン	S_1	利用:防止, 使いすぎ:防ぐ, 誤動作:防ぐ, 置き忘れ:防止, 誤操作:防ぐ, 情報漏えい:防ぐ, いたずら:防止, スマホ, 健康被害:防止, 反射:防止
	S_2	落下リスク:軽減, 負担:軽減, 温度上昇:軽減, ストレス:軽減, 負荷:軽減, 電磁波被ばく:軽減, 電池消耗:軽減, 電磁波:軽減, 初期費用:軽減, 音漏れ:低減
ロードバイク	S_1	日焼け:防止, チェーン脱落:防止, 水:防止, 傷:防止, 汚れ:落とす, 錆:防ぐ, 脱落:防止, 転倒:防ぐ, 振動:防止, ズレ:防ぐ
	S_2	負担:軽減, 痛み:軽減, 疲れ:軽減, 疲労:軽減, 空気抵抗:軽減, 振動:軽減, ロードノイズ:軽減, 両方:軽減, 突き上げ:軽減, ストレス:軽減
ギター	S_1	汚れ:落とす, 塗装:落とす, くすみ:落とす, 色:落とす, チューニング:変える, ピック:落とす, テンポ:落とす, 鳥:変える, サビ:落とす, 傷:防止
	S_2	負担:軽減, ノイズ:軽減, 狂い:軽減, ストレス:軽減, 疲れ:軽減, 摩擦音:軽減, キック振動:軽減, 摩擦:軽減, 力:感じる
掃除	S_1	汚れ:落とす, カビ:落とす, 油汚れ:落とす, 水垢:落とす, ホコリ:掃除, 尿石:落とす, 黒カビ:落とす, 臭い:取る, 黄ばみ:落とす, シミ:落とす
	S_2	手間:かける, 負担:軽減, ムダ:省く, ひと手間:省く, 時間:かける, 無駄:省く, 物:減らす, 臭い:減らす
洗濯	S_1	汚れ:落とす, 臭い:取る, 黄ばみ:落とす, ニオイ:防ぐ, におい:取る, カビ:落とす, シミ:落とす, カビ臭さ:取る, 匂い:消す, 加齢臭:落とす
	S_2	汚れ:落とす, 臭い:取る, 黄ばみ:落とす, ニオイ:取る, カビ:落とす, におい:取る, シミ:落とす, カビ臭さ:取る, 匂い:落とす, エリ汚れ:落とす
登山	S_1	高山病:防ぐ, 事故:防ぐ, 遭難:防ぐ, 道迷い:防ぐ, 凍結:防ぐ, 汗冷え:防ぐ, 道迷い遭難:防ぐ, 遭難事故:防ぐ, 低体温症:防ぐ, 山岳遭難事故:防ぐ
	S_2	ストレス:解消, 運動不足:解消, 痛み:解消, 筋肉痛:解消, 靴ずれ:解消, 浮腫み:解消, 段差:解消, ムレ:解消, 痛さ:解消, 曇り:解消
観光	S_1	侵入:防ぐ, 火災:防ぐ, 事故:防ぐ, 被害:受ける, 熱中症:防ぐ, 交通事故:防ぐ, 侵略:防ぐ, 忘れ:防ぐ, 詐欺:防ぐ
	S_2	ぼったくり被害:補償, 被害:補償, 海外旅行保険:補償, 損害:受ける, ケガ:補償, 期間:観光, 風力発電被害:補償, 時価額:補償
海外旅行	S_1	パケ死:防ぐ, 時差ボケ:防ぐ, 事故:防ぐ, 体重増加:防ぐ, 盗難:防ぐ, ケンカ:防ぐ, 「時差ボケ」:防ぐ, 汚れ:防ぐ, 喧嘩:防ぐ
	S_2	時差ボケ:解消, 「時差ボケ」:解消, 痛み:解消, わずらわしさ:解消, 睡眠時差ボケ:解消, 野菜不足:解消, デフレ:解消, 暗さ:解消, 運動不足:解消, 心配:解消

表 2　シード動詞集合 S_1, S_2 により抽出されたリスク集合

クエリ	シード	抽出された動詞
ヘッドフォン	S_1	防ぐ, 防止, 低減, 解消, 軽減, 解決, 抑える, 除去, 徹底, 拾う
	S_2	低減, カット, 軽減, 消す, 拾う, 減少, 遮断, 除去, 防ぐ, 検知
パソコン	S_1	防止, 防ぐ, 知らせる, 抑止, 監視, カット, 受ける, 特定, 解消, 阻止
	S_2	軽減, 取る, かける, とる, 解消, 減らす, 防ぐ, 癒す, 抑える, 感じる
スマートフォン	S_1	防ぐ, 防止, 防げる, 軽減, 抑える, 帳消し, やめる, 心配, 引き起こす, 懸念
	S_2	軽減, 低減, かける, 抑える, 防止, 検知, 減らす, 抑制, 感じる, 防ぐ
ロードバイク	S_1	防止, 防ぐ, 落とす, 避ける, 経験, 取る, 直す, 考慮, 繰り返す, 引く
	S_2	軽減, かける, 感じる, 抑える, 考える, 減らす, 低減, 解消, 緩和, 和らげる
ギター	S_1	落とす, 取る, 取り除く, 拭き取る, 再現, ふきとる, 除去, 防止, とる, 変える
	S_2	軽減, かける, 減らす, 与える, 抑える, 和らげる, 感じる, 掛ける, 考える, お願い
掃除	S_1	落とす, 掃除, 取る, ためる, 一掃, こする, 除去, 防ぐ, 取り除く, とる
	S_2	かける, 省く, 軽減, 減らす, カット, 削減, 解消, なくす, 省ける, 惜しむ
洗濯	S_1	落とす, 取る, 消す, 防ぐ, とる, おとす, 抑える, 知る, 分解, 予防
	S_2	落とす, 取る, 消す, 防ぐ, とる, おとす, 知る, 抑える, 分解, 予防
登山	S_1	防ぐ, 考える, 回避, 防止, 予防, 発症, 招く, 経験, 辿る, 防げる
	S_2	解消, 感じる, 痛感, 緩和, 軽減, 癒す, 発散, 実感, 減らす, 思い知らす
観光	S_1	防ぐ, 防止, 予防, 受ける, 想定, 許す, 発見, なくす, 発する, 遮る
	S_2	補償, 受ける, 観光, 防ぐ, 申告, 懸念, 払拭, 訴える, もたらす, 吹き飛ばす
海外旅行	S_1	防ぐ, 防止, 解消, 治す, 経験, 除く, 恐れる, 予防, 補償, 考える
	S_2	解消, 防ぐ, 感じる, 治す, 補う, 軽減, 防止, 補える, 考える, 経験

表 3　シード動詞集合 S_1, S_2 により抽出された動詞

とす」と「おとす」という 2 つの表記が異なるだけの動詞が含まれてしまっているし,「掃除」というクエリについても「取る」と「とる」という表記が異なるだけの動詞が含まれてしまっている. これら表記揺れの動詞が複数含まれていると, その動詞と共起度の高いリスクのスコアが高くなりすぎてしまうという問題が発生してしまう. そのため, これら表記揺れをまとめて 1 つの語として扱う必要があると考えられる. 表記揺れを同じ語として扱うための手法としては, MeCab で得られる形態素の読みの情報を利用する手段が考えられる.「洗濯」というクエリでの抽出の際に問題となった「落とす」と「おとす」という 2 つの動詞に関する問題だと, これら 2 つの形態素における読みは「おとす」で一致するため, まとめることが可能となる. ただし, MeCab の読みを利用するだけでは, 内容が同じ動詞が複数含まれてしまう問題を解決しきることはできない.「防止」という動詞と「防ぐ」という動詞は意味的には同じで抽出されるリスクも類似性が高いものとなるが, これらの動詞は MeCab の読みを利用する方法ではまとめることができない. これらの動詞を同一のものとして扱う手法としては同意語の辞書を利用する方法が考えられる. しかし, 辞書を利用する方法でもまとめきれない動詞も出現すると考えられるので, 多様な動詞の抽出に重点をおいたスコア付けの手法を新たに追加する必要がある可能性も示唆される.

リスク集合や動詞集合を求める際に, 上位 10 個に絞ってしまっていた事も多様性の低下に大きく関係していると思われる. リスク集合や動詞集合の大きさを制限した理由としては, 検索 API の利用回数に上限があり無駄なアクセスを減らす必要があったことが挙げられるが, 10 個に絞ってしまったことにより, 多様な動詞の取得を行いたいという目的が阻害されてしまったように思われる.

本研究の抽出アルゴリズムではリスクを抽出する際に動詞のスコアに加えて, 前回抽出されたリスクを用いた動詞の適格性というスコアを用いていた. 動詞を抽出する際には, リスクのスコアに加えて, 前回抽出された動詞を用いたリスクの適格性というスコアを用いていた. つまり, 動詞の抽出部分とリスクの抽出部分で同じアルゴリズムを利用していたことになる. このことが, 抽出精度や多様性の低下に影響した点もあると考えられる. あるクエリに関して抽出するべきリスクは無数にあるが, そのリスクを抽出し, 評価するのに有用な動詞の数はリスクの数と比べてかなり少ないものになると考えられる. そのため, リスク集合に含まれるリスク数の最大値をもっと大きな数にしておく必要性が懸念される. 実際, 多くのクエリに関して, スコアが最も高い動詞によって抽出されたリスクが取得されたリスク集合の大半を占めていたことからも, リスク集合の大きさによる問題があったことが伺える. このようにリスク集合取得部分と動詞集合取得部分のアルゴリズムをそれぞれの特徴や目的に合わせて最適化する処置が今後必要になると考えている.

6. ま と め

本論文では Web 上のリスクへの対策が記述されている文書を利用してリスクを抽出し, 発生頻度・影響度という観点で評

価する手法を提案した.

本論文で達成できなかったこととしては, ポジティブな状態に移行するリスクを抽出・評価することができなかったことが挙げられる. リスクにはダウンサイドリスクとアップサイドリスクがあり, 本論文で取り扱ったのはダウンサイドリスクである. ダメージを受けた状態に移行しないようにするための情報に価値があるのと同様に, 利益を得られる状態に移行するための情報にも価値があると考えられる. 本論文で提示した手法を改良し, ダウンサイドリスクとアップサイドリスクの両方に対応できるようにすることを今後の課題として考えている.

また, 動詞を用いてリスクを分類し評価する部分の実装ができていないことは今後最も早く解決すべき課題であると考えている. 「ヘッドフォン」というクエリにおける理想の分類例を図 2 に示す. 本研究では現状リスクの抽出とそれぞれのリスクの抽出に有用な動詞の取得まではできているが, 動詞を利用したリスクの分類についてはまだ実装まで至っておらず, 課題が残っている.

また, 取得されたリスク集合の多様性の低さを解消することも今後の課題であると言える. 抽出された動詞集合はリスク集合に比べると多様性があるといえるが, 動詞の抽出に用いた手法について, 多様な動詞を取得するための手法としては不十分な点が見受けられるため, 改善が必要である. 現状のブートストラップ法による抽出で利用するスコアは, スコアが高いリスクに 1 つでもリンクしていると動詞のスコアが高くなるというものになっている. このような手法を用いた場合, あるスコアの高い 1 つのリスクにリンクしている動詞が数多く取得されることが多くなると考えられ, 実際の実験結果からもその傾向が伺えた. そのため, 動詞がリンクしているリスクの数がスコアに与える影響が多くなるようにスコア付けの方法を変更することで, リスク抽出に有用かつ多様な動詞を含む動詞集合の取得ができるようになる可能性があると考えられる.

本研究の最終的な目的はユーザが入力したクエリに関わるリスクを抽出し, 発生頻度・影響度という観点で評価するための手法を提案することである. この論文ではリスクの抽出に関してリスクを目的語にとる動詞が有用であるという仮説をたて, その仮説に基づき多様なリスクを抽出するための手法を提案し, また, リスクと動詞の関係を利用して, リスクを 4 つの対策パターンに分類するための手法を提案した. 実際に実験を行った知見として, 動詞によりある程度のリスクを抽出することができることは照明できたが, 動詞の情報だけでは発生頻度・影響度を評価するのに十分であるかどうかについては明確な解答が出せていなかった. 今後は発生頻度・影響度の推定のために有用な動詞以外のファクターについても考慮を重ね, 幅広いドメインにおけるリスクの評価を行えるようにしたいと考えている.

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.

図 2　動詞を利用したリスクの分類例

文　　献

[1] Kentaro Inui, Shuya Abe, Kazuo Hara, Hiraku Morita, Chitose Sao, Megumi Eguchi, Asuka Sumida, Koji Murakami, and Suguru Matsuyoshi. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. In *WI-IAT '08 Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, 2008.

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

[3] Robert J. Walker, Reid Holmes, Ian Hedgeland, Puneet Kapur, and Andrew Smith. A lightweight approach to technical risk estimation via probabilistic impact analysis. In *MSR '06 Proceedings of the 2006 international workshop on Mining software repositories*, 2006.