高精度の WWW&パーソナルサーチエンジン
— あなたのvocabulary(こ と ば)であらゆる情報にアクセス！ —

1. **背景**

   開発者は平成 14 年度未踏ソフトウェア創造事業「多様なトピックに対応する高精度の情報検索システム」において，情報検索のための「関連性の重ね合わせモデル (Relevance-based Superimposition Model; **以下** RS **モデル**) **を多様なデータベースに対応させるにあたり重要であった，トピックを表す語句を予め人間が付与しておく作業を自動化する自動重要語抽出技術を開発し，検索精度向上における高い効果を示した．**

   その後，「多様な〜」で開発を行なった検索システムの， World Wide Web(WWW) **上の** HTML **文書等の大規模な文書集合や，メーリングリスト等の比較的小規模な文書集合に対する特性を検証した．その結果，** WWW **ではトピック抽出精度が十分でない場合があること，小規模文書集合ではトピック抽出精度は十分なもののトピックに対して補うべき語彙が小規模文書集合自身からは十分に獲得できない場合があることが問題点として認識された．**

   本プロジェクトではこの二つの問題を解決して検索精度を向上させ，適用分野の拡大を図った．

2. **目的**

   開発者は平成 14 **年度未踏ソフトウェア創造事業において，** RS **モデルに基づく情報検索システムを開発した．本プロジェクトでは，適用分野の拡大と更なる高性能化を目指し，以下の機能を持ったシステムの開発を行った．**

   - WWW **上の情報をリンク解析，言語的解析を駆使して分類し，** RS **モデルの適用に最適な情報粒度に再構成するシステムの開発**
   - WWW **から語彙を獲得し，個人利用者や企業内データベースなど小規模な文書集合に対して適切な索引語を補うシステムの開発**

3. **開発の内容**

   本プロジェクトは以下の機能を開発した．

   (a) WWW **サーチエンジンのための，ハイパーテキストのクラスタリング**

   (b) **パーソナル文書検索のための，小規模文書集合の非排他的クラスタリング**

   (c) **パーソナル文書検索のための，小規模文書集合の文書クラスタと** WWW **文書クラスタの特徴ベクトル重ね合わせ**

4. **従来の技術との相違**

情報検索においてクラスタリングを応用する技術としては Scatter / Gather のように検索結果を動的に分類して整理する技術などが開発されているが，本プロジェクトではリンク構造解析によるクラスタリング技術を検索精度の向上のために応用する点が新規的である．

RS モデルについても，平成 14 年度にはニュース記事検索での有効性を示したのに続き，WWW に対する有効性と，小規模文書集合に対する有効性を示した．検索システムの評価トレンドは大規模化に向かっているが，その先には個人が蓄積した情報に対する検索要求が高まってくると予想し，小規模文書集合での検索精度向上にも敢えて挑戦したものである．

表 1: パーソナル文書検索の性能（平均適合率）

※ 1: 正解セグメント数 / 全セグメント数．セグメントとは文書をトピックで自動分割したもの．
「RS モデル」の括弧内の数値は Namazu を 1 とした場合の向上率．
「RS+WWW 文書の重ね合わせ」の括弧内の数値は順に，Namazu を 1 とした場合と「RS モデル」を 1 とした場合の向上率．

問い合わせ	※ 1	Namazu	RS モデル		RS+WWW 文書の重ね合わせ	
「松井秀喜」	57 / 866	0.7083	0.6914	(−2.4%)	0.7136	(+0.7% / +3.2%)
「野球選手松井秀喜」		0.4681	0.6855	(+46.4%)	0.7098	(+51.6% / +3.5%)
「ゴジラ松井」		0.5972	0.5916	(−0.9%)	0.6147	(+2.9% / +3.9%)
「ゴジラ」		0.0925	0.3919	(+323.7%)	0.4654	(+403.1% / +18.8%)
「松井稼頭央」	9 / 866	0.7965	0.5674	(−28.8%)	0.6119	(−23.2% / +7.8%)
「野球選手松井稼頭央」		0.4201	0.5674	(+35.1%)	0.6484	(+53.9% / +14.3%)
「リトル松井」		0.3812	0.2519	(−33.9%)	0.2515	(−34.0% / −0.2%)
「リトル」		0.2222	0.2222	(± 0%)	0.2222	(± 0% / ± 0%)
「人材派遣 (にどのような) 業種 (があるのか)」	153 / 1659	0.8680	0.7693	(−11.4%)	0.7991	(−7.9% / +3.9%)
「日本全国 (の) ショッピングモール」	4 / 397	0.0314	0.0310	(−1.3%)	0.0509	(+62.1% / +64.2%)

5. **期待される効果**

RS モデルを検索システムに導入することにより，自然な問い合わせ表現で高精度の検索を行なうことが可能となる．特に，個人・小規模グループが蓄積した文書の検索システムでは，これらの小規模文書集合に最適化された技術を搭載することで検索洩れの低減を達成した．検索対象の文書以外から語彙を自動的に補強する機能は，検索対象が小規模な文書集合以外でも有効であるし，語彙を提供する文書群も問い合わせに用いられる語彙を網羅するものであれば WWW に限定されない．

以下に，本開発成果の応用可能性を幾つか述べる．

- **企業のサポート業務の効率化**
  企業のヘルプデスク業務，あるいは WWW の「よくある質問」ページの利用者の中には専門用語，企業特有の言い回しが障害となって十分な

情報を得られないという経験を持つ人が少なくない．WWW 全体からの語彙補強を行ない，企業側の情報を平易な表現で検索可能にしておくことにより，このような問題を解消できると考えられる．

- **企業間，部門間の情報共有を促進**

  企業間，部門間には大きな語彙の差が存在しており，コラボレーションを阻害しているケースが多い．この場合，共有した情報に対してそれぞれの企業，部門がそれぞれの視点で語彙補強を行なうことにより，常用している語彙での情報取得が可能となる．

6. **普及・活用の見通し**

   個人向け，イントラネット向け情報ツールとしてケイ・ワイ・エイ・グループからの製品リリースを予定している．

7. **開発者**: **金沢 輝一** (**ケイ・ワイ・エイ・グループ**, tkana@kyagroup.com)