2C2-02

インターネットリソースを活用した情報技術マップ構築支援環境

Automatic Generation of the Multiple-phased Technology Maps from the Internet Resources

小林 慎一[*1] 白井 康之[*1] 粂野 文洋[*1] 犬島 浩[*2]
Shin-ichi Kobayashi Yasuyuki Shirai Fumihiro Kumeno Hiroshi Inujima

[*1](株) 三菱総合研究所 Mitsubishi Research Institute, Inc.
[*2]早稲田大学大学院 情報生産システム研究科 Waseda University

Information technology is increasingly crucial in recent years for the development of our society. IT has brought many changes to everything in our society with incredible speed. Hence, when we investigate R & D themes or plan business strategies in IT, we must understand overall situation around the target technology area, such as products, organizations, standards and markets besides technology itself. For this purpose, we developed a method to generate Multiple-phased trend maps automatically based on the Internet content. Furthermore, we introduced quantitative structural indicators to evaluate maps. According to our evaluation of this method we got succesful and interesting results.

1. はじめに

近年の社会・産業の特徴は，技術がそれらと極めて密接に関連していることである．とりわけ情報技術は人間の知的活動全般を支援しその範囲を拡大するものであることから，多くの産業や社会全般に横断的に関与する基盤的学問分野となっている．さらに情報技術は関連する分野が広いだけでなく，その変化のスピードが急速に加速されていることも併せて指摘できる．この結果，情報技術分野の新規の研究開発テーマや事業展開の可能性を検討し判断するためには，対象として想定している分野に関して単に技術に限定しない総合的な状況把握が必要となる．すなわち，技術に関連する組織，規格，サービス，社会現象/トレンドなどの複数の視点からの情報を相互に関連付けて（本研究ではこれを多相な情報と言う）総合的な理解を得なければならない．加えて，変化のスピードに対応するためにはインターネットに流通する優良なニュースサイト情報などの良質なリソースのリアルタイム的体系化および時系列的な変化の認識も必須である．本研究では，このような背景のもと，インターネット情報を用いて，技術的概念を総合的に理解することを支援する技術動向多相マップの自動生成手法に関する提案を行なう．

2. 技術動向多相マップ生成

本システムの構成を 図 1 に示す．インターネットから取得された情報リソースから，パターン定義に従ったキーワード抽出とカテゴリ分類が行なわれる．リソースの解析では，このようにして得られた辞書をもとにして，単語文書行列を生成する．ユーザは，GUI・可視化モジュールを通じて，あるキーワードに近い概念（近傍）を表示させたり，指定されたリソース集合における重要なキーワード間の関連をグラフ化し，その特徴量を把握することができる．

2.1 辞書生成モジュール

正規表現を含むパターンマッチングによってリソースが含むキーワードとそのカテゴリを抽出し辞書として蓄える．パターンは，一般にリソースの種類に依存するため，これらをユーザが定義し，リソース単位でパターンを指定することが可能である．このパターンマッチングによって，9 種類のカテゴリー（相），すなわち，組織名称，技術用語，システム名称，規格名称，サービス名称，人名，ハードウェア名称，社会現象/トレンド，ビジネス用語の各分類に対応したキーワードの自動抽出を可能としている．

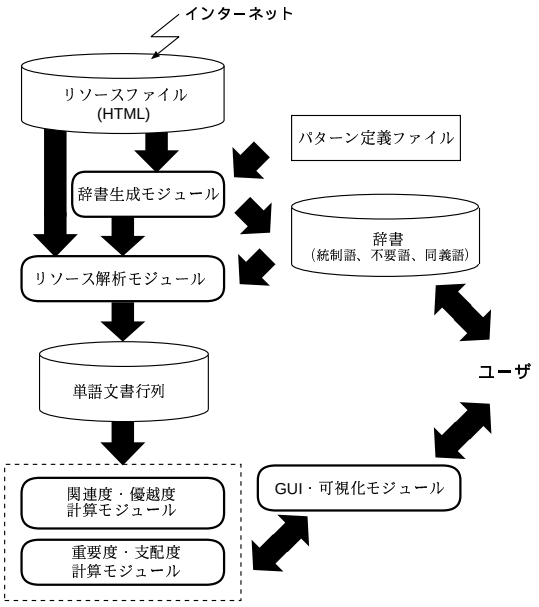

図 1: システムの全体構成

2.2 リソース解析モジュール

辞書に含まれるキーワードが各リソースに出現するか否かを解析し，単語文書行列を生成する．単語文書行列 M は，キーワード数を m，文書数を n とするとき，$m \times n$ 行列であり，各要素 f_{ij} は，キーワード i が文書 j に含まれるとき 1，そうでないとき 0 とする．

2.3 関連度・優越度

関連度とは，2 つのキーワード間において定義され，与えられたリソース集合の中での意味的関連性を示す指標である．ある 2 つのキーワード i と j の関連度 $R(i,j)$ は，以下のように単語ベクトル $\mathbf{t}_i$ と $\mathbf{t}_j$ の余弦で定義する．

連絡先: 小林慎一，(株) 三菱総合研究所 情報環境研究本部，kobayash@mri.co.jp

$$R(i,j) = \frac{\mathbf{t}_i \cdot \mathbf{t}_j}{\|\mathbf{t}_i\| \cdot \|\mathbf{t}_j\|}$$

また，優越度とは，2 つのキーワード間において定義され，与えられたリソース集合の中での意味的な上下関係の抽出を目的としている．i の j に対する優越度 $P(i,j)$ を以下のように定義する．

$$P(i,j) = \arccos(R(i,j^c)) - \arccos(R(i^c,j))$$

2.4 重要度・支配度

上記の関連度，優越度指標は，2 つのキーワードの関係を表す指標であった．以下に示す 2 つの指標は，これに対して，ある指定されたリソース内におけるキーワードの位置付けを示す指標である．

重要度とは，指定されたリソース内における相対的重要性を表す指標であり，キーワード i のリソース集合 R （文書数 n）における重要度指標 $W_R(i)$ を以下のように定義する．

$$W_R(i) = tf_R(i) \cdot \left(log \frac{n}{df_R(i)} + 1 \right)$$

あるリソース集合 R におけるキーワード群 A が与えられたとする．このとき，キーワード $i \in A$ の A に対する支配度 $I_A(i)$ とは，i が A においてどの程度の影響力をもっているかを表すものであり，i が文書中に出現することが，A の他のキーワード $(A \setminus \{i\})$ の出現に対して，どの程度の期待情報量の削減をもたらすかという観点から評価を行なう．すなわち，$i \in A$ の支配度 $I_A(i)$ を以下のように期待情報量縮減比として定義する．

$$I_A(i) = \frac{H(A \setminus \{i\}) - H(A \setminus \{i\} \mid i)}{H(A \setminus \{i\})}$$

支配度は特定のキーワードが当該領域に与える影響の程度を示唆するが，これをカテゴリ単位で考察することによって当該領域における支配的カテゴリを推定することが可能となる．

2.5 GUI・可視化モジュール

GUI は，ユーザからの指定により関連するキーワードをインタラクティブに計算させたり，領域で重要なキーワードを表示させ，各種指標を計算させるためのインタフェースを提供するものである．グラフの可視化部分（自動レイアウト機能を含む）に関しては，Tubingen 大学のプロジェクトに端を発するグラフ描画ライブラリ yFiles [*2] を利用して開発を行なった．図 2 にグラフの表示例を示す．各ノードは，カテゴリ別に色分けがなされている．

3. 実行例

IT 系のニュースメディアである ITmedia [*3] の 2001 年〜2003 年の「ユビキタス」をキーワードとして含む記事を用いた実行例を示す．

3.1 辞書生成

ITmedia に頻出するパターンから，単語切り出し表現を 24 個用意し，辞書生成を行なった．パターンマッチングによる切り出しは，一部不要な切り出しを行なうものもあるが，例えば，連語キーワードの抽出において適切な単位で切り出しができるなどの利点を確認することができた．

*2 by yWorks GmbH (http://www.yworks.de/index.htm)
*3 http://www.itmedia.co.jp

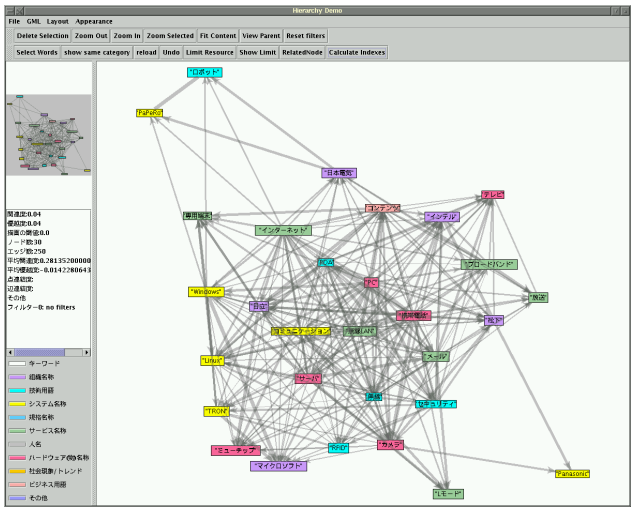

図 2: 「ユビキタス」に関連するキーワードのグラフ表示

3.2 重要度・支配度指標に関する結果

2001 年〜2002 年および 2003 年の記事に対してそれぞれ実行することにより，ユビキタス分野における 2001 年〜2003 年にかけての動向変化を読み取ることができた．例えば，「TRON」は 2002 年までは重要キーワードではなかったが，2003 年には最も重要なキーワードとなっている．ただし，支配度が小さく，このことは，TRON は出現に文脈があって，他の重要キーワードとは独立に出現している傾向があることを示唆する．逆に，重要度に比較して支配度が大きい「ロボット」は，ユビキタス分野の典型的なキーワードではないが他のキーワードへの影響が大きく，いわば間接的・媒介的な意味合いでユビキタス文脈の中で重要な位置を占めていると解釈できる．

4. まとめと今後の課題

より実用的な技術マップを構築することを目的として，キーワードのカテゴリ化と関連度，優越度，重要度，支配度の各指標を導入した技術マップ構築環境について述べた．本手法は，IT 技術マップのみならず，人脈マップ，製品マップ，特許マップ等，さまざまな領域への拡張が可能である．

今後の課題としては，以下が挙げられる．まず第一に，本システムが対象とするリソースは膨大な量にのぼり，これを管理し，各種指標を計算させるには，より効率的なデータ管理方法，指標計算方法が必要とされる．第二に，本システムを現場において実際に活用するためには，データを定期的にかつ自動的に取得し，解析を行なった上でデータを統合する仕組みが必要である．第三に技術動向多相マップの定量的評価をより精緻化させるために新たな指標として相互関連性指標，発展性指標，波及性指標などを導入したいと考えている．

謝辞

実験に使用した ITmedia の各コンテンツの使用を快く許諾して頂いたソフトバンク・アイティメディア（株）に感謝致します．