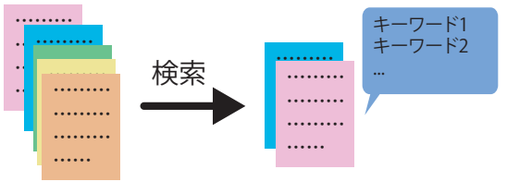
# 単語と文書間における複数の関係性を用いたキーワード抽出

東京都立産業技術高等専門学校　横井　健

## 概要

文書検索における上位にランキングされた文書群
→ キーワードの抽出

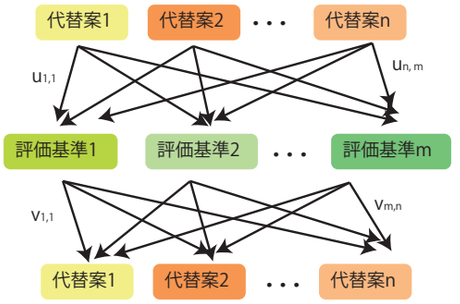



複数の基準を最適化したキーワード抽出手法
・文書中における単語の出現頻度
・単語同士の共起関係
・文書同士の類似性

上記の基準をAnalytic Network Processing (ANP) を拡張した最適化手法で最適化
→ 各単語に付与される重みでキーワードを決定

## 要素技術

1. ANP: Analytic Network Processing (Satty. T. L., 1996)
ANP
・意志決定問題における統一的な評価を決定
・文書は「単語をどのように配置するか」という意志決定問題
ANPでは代替案と評価基準の相互評価を最適化
→ 本研究では代替案：単語、評価基準：文書
相互評価を重みとするネットワーク構造を構築
→ 超行列Sにより表現




$$S = \begin{bmatrix} 0 & V \\ U & 0 \end{bmatrix}$$

V: 単語 → 文書（文書数×単語数）
U: 文書 → 単語（単語数×文書数）

超行列Sの主固有ベクトル
→ 総合評価（文書の評価＋単語の評価）

2. 提案手法
ANPの超行列**S**を拡張
→ 単語間、文書間の特徴を導入した$\mathbf{S}_{ext}$

$$S_{ext} = \begin{bmatrix} D & V \\ U & W \end{bmatrix}$$

D: 文書間の特徴
W: 単語間の特徴

U, V, D, Wの構成
U (文書が与えられたときの単語の評価) の構成
→ ある文書における単語の出現確率で表現

$\mathbf{u}_i = \left[p(w_1|\mathbf{d}_i) \quad p(w_2|\mathbf{d}_i) \quad \cdots \quad p(w_n|\mathbf{d}_i)\right]^T$ *n*: 文書郡中の全単語数

$U = \left[\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_m\right]$ *m*: 文書郡中の文書数

$p(w_j|\mathbf{d}_i) = tf_{ji} / \sum_{j=1}^{n} tf_{ji}$ $tf_{ji}$:文書*i*中の単語*j*の出現頻度

V（単語が与えられたときの文書の評価）の構成
→ ある単語が与えられたときの文書の出現確率

$\mathbf{v}_j = \left[p(\mathbf{d}_1|w_j) \quad p(\mathbf{d}_2|w_j) \quad \cdots \quad p(\mathbf{d}_m|w_j)\right]^T$

$V = \left[\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_n\right]$

$p(\mathbf{d}_i|w_j)$ は以下のようにベイズの定理により導出

$$p(\mathbf{d}_i|w_j) = \frac{p(w_j|\mathbf{d}_i)p(\mathbf{d}_i)}{p(w_j)} \qquad p(\mathbf{d}_i) = 1/m$$

$$p(w_j) = \sum_{i=1}^{m} p(w_j|\mathbf{d}_i)p(\mathbf{d}_i)$$

W（単語間の特徴）の構成
→ 単語*i*の全出現頻度に対する単語*j*との共起頻度

$$W_{ij} = \frac{co(i,j)}{\sum_{j=1}^{n} co(i,j)}$$

*co(i,j)*：単語*i*と単語*j*の共起頻度

・共起：同一文書内に出現すること
・*i* = *j* のときは $\mathbf{W}_{ij} = 1$

D（文書間の特徴）の構成
→ 文書*i*と*j*の正規化類似度

$$D_{ij} = \frac{sim(\mathbf{d}_i, \mathbf{d}_j)}{\sum_{j=1}^{m} sim(\mathbf{d}_i, \mathbf{d}_j)}$$

$sim(\mathbf{d}_i, \mathbf{d}_j)$：cos尺度、dice係数、jacard係数

## 適用実験

1. 実験に用いた検索結果文書群
2009年6月1日に*Yahoo!*検索エンジンで"finance", "pandemic" の2つの単語で検索した英語の検索結果上位50件

**2.** 比較手法
各単語のtf-idf値、df値（文書頻度）、tf値（単語頻度）の文書群における総和

3. 実験結果

表1. "pandemic"の検索結果における重要語

| 手法 | 重要度ランキング上位10単語 |
|---|---|
| ANP | pandem influenza flu plan legion crazymonkeygamescom occur ar thi outbreak |
| ExtANP(cos) | pandem influenza mere becaus diseas flu infecti definit condit contagi |
| ExtANP(dice) | pandem influenza mere becaus flu diseas infecti definit condit contagi |
| ExtANP(jac) | pandem influenza mere becaus flu diseas infecti definit condit contagi |
| tf-idf | flu influenza plan ar legion thi involv occur prepared outbreak |
| df | pandem influenza flu plan thi occur prepared outbreak diseas health |
| tf | pandem influenza flu plan ar thi occur prepared outbreak diseas |

表2. "finance"の検索結果における重要語

| 手法 | 重要度ランキング上位10単語 |
|---|---|
| ANP | financialcrisi tag crisi financi post thi photo rate nation comment |
| ExtANP(cos) | financialcrisi tag crisi financi post thi photo rate comment video |
| ExtANP(dice) | financialcrisi tag crisi financi post thi photo rate comment video |
| ExtANP(jac) | financialcrisi tag crisi financi post thi photo rate comment video |
| tf-idf | tag financialcrisi financi crisi post thi comment photo rate nation |
| df | financialcrisi tag financi crisi post photo thi rate nation global |
| tf | financialcrisi tag financi crisi post thi photo comment rate nation |

## まとめ

・複数の基準を最適化したキーワード抽出手法
→ 従来のdfやtfによるキーワードに近いものを抽出
提案手法の方が有用性が高いキーワードも含んでいる傾向
・今後の課題
現状では文書の出現確率は一様
→ ユーザの嗜好や検索結果順位の考慮
文書そのものへも重み付けが可能