2005－DBS－137（Ｉ）（14）
2005／7／13

汎用アノテーションシステム（MAML System）を利用したWeb検索結果のグラフ表示

滝本　湖[†]　伊藤　一成[††]　斎藤　博昭[†]

† 慶應義塾大学 大学院理工学研究科
〒 223–8522 神奈川県横浜市港北区日吉 3–14–1
†† 青山学院大学 理工学部
〒 229–8558 神奈川県相模原市淵野辺 5–10–1
E-mail: †{lake, hxs}@nak.ics.keio.ac.jp, ††kaz@it.aoyama.ac.jp

あらまし　Web 検索において，サーチエンジンの利用が一般的になっている．しかし，サーチエンジンが返す検索結果数は膨大であり，単にランク付けされたリストからユーザが望みの Web ページを見つけ出すのは困難である．そこで，本稿では我々が既に提案している汎用アノテーションシステム MAML System を用いて，Web 検索結果を効果的に提示する手法を提案する．まず，内容に基づくクラスタリングにより Web 検索結果をグループ化する．また，類似度が高い文書同士を二次元上の近傍に配置する．これにより，ユーザは直感的かつ大局的に Web 検索結果を把握できる．

キーワード　Web 利用技術, 情報検索, テキストマイニング

Graphical Presentation of Web Search Results Using MAML System

Hiroshi TAKIMOTO[†], Kazunari ITO[††], and Hiroaki SAITO[†]

† Graduate School of Science and Technology, Keio University
Hiyoshi 3–14–1, Kouhoku–ku, Yokohama, Kanagawa, 223–8522 Japan
†† College of Science and Engineering, Aoyama Gakuin University
Fuchinobe 5–10–1, Sagamihara, Kanagawa, 229–8558 Japan
E-mail: †{lake, hxs}@nak.ics.keio.ac.jp, ††kaz@it.aoyama.ac.jp

Abstract　While using search engines is a popular way of Web browsing, it is difficult for users to get Web pages of interest from a huge amount of search results returned as ranked lists. This paper proposes an efficient way of presenting search results by using MAML System. We group similar Web pages together and visualize those clusterd search results as a graph. Our proposed method facilitates users assessing search results intuitively and in perspective.

Key words　web technology, information retrieval, text mining

1. はじめに

近年のインターネットの普及に伴い，Web ページの閲覧は日常的な行動となっている．その際，望みの Web ページを見つけるために Web 検索を行う機会が増えている．しかし，サーチエンジンによる検索においては，提示される検索結果件数はしばしば膨大なものとなる．このような場合，ユーザがすべての検索結果の Web ページを閲覧することは事実上不可能であり，結果として有用である情報を見逃してしまう可能性がある．また，検索結

果はランク付けされたリスト形式で表示されるが，これを順番に精査していく作業は非生産的といわざる得ない．

ここで，以上の問題に対して以下のようなアプローチで改善を試みる．第一に，Web ページ群に対するクラスタリング処理が考えられる．あらかじめグループ分け，取捨選択した Web ページを提示することにより，検索結果数が膨大であってもユーザが効率的に Web 閲覧を行えるようになることが期待できる．第二に，Web ページ群のグラフ視覚化が考えられる．グラフ上に視覚化された情報を閲覧することにより，ユーザが検索結果を直感的かつ大局的に把握・利用可能になることが期待できる．

本稿では，上記 2 種類のアプローチを併用して， Web 検索結果のクラスタリングとその視覚化を行う手法を提案し，Web 検索結果の閲覧性向上を試みる．

2. 関 連 研 究

Web 検索結果のクラスタリングに関する試みとしては，Zamir らが Suffix Tree Clustering(STC) という手法を提案している [1]．STC では，文書中に出現する単語列によって形成されるトライを生成することで文書集合中に含まれる高頻度の単語列を効率良く見つけることができる．Zamir らはさらに，上記の手法を利用したメタサーチエンジン Grouper を提案している [2]．Kummamuru らは，Web 検索結果の階層的クラスタリング手法を提案している [3]．この手法は単語ベースのクラスタリングであり，各単語に対して文書網羅性とトピック分離性という 2 種類の基準を用いてスコアを計算する．

一方，Web 検索結果の視覚化に関する試みとしては，TouchGraph 社が開発した GoogleBrowser というシステムが挙げられる [4]．GoogleBrowser の主な機能は，入力された Web ページの URL を元に Google の関連ページ検索を繰り返し行い，関連のあるページを相互にエッジで結合したグラフとして表示するというものである．anacubis 社は Google-enabled visual search というシステムを開発した [5]．これは Google の関連ページ検索とリンクページ検索の結果をグラフ表示するものである．

3. Web 検索結果のクラスタリング手法

本提案で用いるクラスタリング手法のステップを順に示す．

3.1 単語集合の抽出

はじめに，クラスタリング対象の文書集合から文書内に出現する単語を抽出する．

3.1.1 名詞・未知語の抽出

文書内に出現する語のうち，動詞，接続詞，助詞といった品詞の語は，その文書を表現する語としてはあまり有効ではない．そこで，品詞が名詞，未知語である語を抽出対象とする．

3.1.2 出現頻度が過多・過少な語の除去

単語を抽出する際，文書ごとの出現頻度と文書集合全体の出現頻度の二種類を求める．文書集合全体の出現頻度が過多な語は，どの文書にも出現する傾向があるため文書を特徴付ける効果が少ない．また出現頻度が過少な語は，文書を特徴付ける働きはあると考えられるが，文書集合全体への影響が少なく文書をグループ分けする際にはあまり有効ではない．そこで，出現頻度が閾値 th_{high} より高い，あるいは閾値 th_{low} より低い語は除く．この結果抽出された単語集合を $T = (t_1, t_2, ..., t_m)$ とする．

3.2 単語ベクトルの生成

抽出された単語集合 T を基に，文書の単語ベクトルを生成する．文書 d_i の単語ベクトル $\vec{v}(d_i)$ は次のように定義する．

$$\vec{v}(d_i) = (w_1, w_2, ..., w_m) \quad (1)$$

$$w_k = TF(t_k, d_i) \quad (2)$$

ここで，w_k は d_i における語 t_k の重み，$TF(t_k, d_i)$ は d_i における語 t_k の出現頻度とする．

3.3 クラスタの生成

3.3.1 初期クラスタの生成

各文書をそれぞれ一つのクラスタとみなし，初期クラスタとして生成する．クラスタを C_i，文書を d_i，全文書数を n とすると，初期クラスタは次のようになる．

$$C_1 = \{d_1\}, C_2 = \{d_2\}, ..., C_n = \{d_n\} \quad (3)$$

3.3.2 クラスタの統合

クラスタ C_i の単語ベクトル $\vec{v}(C_i)$ を，そのクラスタに含まれる文書の各単語ベクトルの重心として次のように定義する．

$$\vec{v}(C_i) = \frac{\sum_{d \in C_i} \vec{v}(d_i)}{|C_i|} \quad (4)$$

クラスタ間の類似度は，コサイン類似度を用いて以下のように定義する．

$$sim(C_i, C_j) = \frac{\vec{v}(C_i) \cdot \vec{v}(C_j)}{|\vec{v}(C_i)||\vec{v}(C_j)|} \quad (5)$$

以下，クラスタの統合手順を示す．クラスタ間の類似度を比較し，最も類似度の高い２つのクラスタを１つのクラスタに統合する．統合するクラスタを C_i，C_j とすると，統合後のクラスタ C_{new} は以下のように定義される．

$$(i, j) = \arg\max_{i,j} sim(C_i, C_j) \quad (6)$$

$$C_{new} = C_i \cup C_j \quad (7)$$

クラスタの統合は，クラスタ間の最大類似度が閾値 th_{sim} より小さくなるまで繰り返し行う．

3.3.3 クラスタの選別

クラスタの統合が終了したら，統合されたクラスタ集合の中から最終的に作成するクラスタを選別する．統合されたクラスタ集合を C_{merge}，最終的に作成するクラスタ集合を C_{final} と定義する．C_{merge} において，クラスタに含まれる文書数が閾値 th_{elem} より大きいクラスタはそのまま C_{final} の要素とする．クラスタに含まれる文書数が th_{elem} 以下のクラスタは，そのクラスタをクラスタ集合 C_{other} に統合し，選別処理の最後に C_{other} を C_{final} に追加する．

3.4 ラベルの生成

クラスタのラベルを生成する．まず，クラスタに属している各文書の単語ベクトルの和を求める．求めた和ベクトルの要素の中から重みが最も大きい上位 2 つの単語を，そのクラスタのラベルとして抽出する．なお，クラスタ C_{other} のラベルは「その他」とする．

4. クラスタリングと視覚化を用いた Web 検索結果の提示手法

3 章で提案した手法を利用した，Web 検索結果のクラスタリングと視覚化を行うシステムについて述べる．

本システムを構築するにあたり，我々が提案する汎用アノテーションシステム MAML System [7] を利用した．以下，MAML System の概要と本システムの詳細を説明する．

4.1 MAML System の概要

MAML System とは，我々が提案するアノテーション記述言語 MAML（Multimedia Annotation Markup Language）[6] により記述された情報を処理するインタフェースである．MAML はメディアの種類やフォーマットに依存しない統一的な記述仕様を用いたアノテーション記述言語であり，本稿で扱う Web 検索結果情報も容易に記述可能である．

MAML System は，ユーザが二次元空間上に様々な情報とそれに対する注釈情報（アノテーションデータ）を配置し，それらの情報を管理・活用できる．最大の特徴は，アノテーションを生成するプロセスと利用・処理するプロセスが同一インタフェース上で実現されるだけではなく，それに係る一連の操作まで全く同じ点である．ユーザがアノテーションを生成することと何らかの処理を行うことは，作成するアノテーションの種別が異なるだけである．さらに二次元空間上に配置された情報に対し，マウスを用いた直感的な操作が可能なインタフェースを提供している．

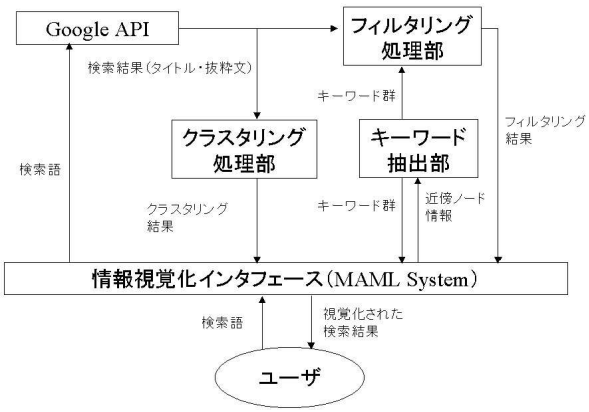

図 1　システム構成図

4.2 システム構成

本システムの構成を図 1 に示す．

4.2.1 情報視覚化インタフェース

MAML System を用い情報を二次元空間上に配置して画面上に表示する機能を提供する．ユーザは Web 検索を行う際，インタフェース上で Web 検索用のアノテーションノード（以下 Web 検索ノードと呼ぶ．）を作成し，そこに検索語を入力する．検索語が入力されると，Google API [8] に検索処理が委託される．

Google API からは検索結果として Web ページの URL，タイトル，抜粋文が返され，それらがクラスタリング処理部とフィルタリング処理部へと渡される．またキーワード抽出前に，検索を行う Web 検索ノードの近傍に存在するノードの情報（ノード内のテキストと Web 検索ノードまでの距離）がキーワード抽出部へと渡されて利用される．クラスタリング処理部とフィルタリング処理部からはそれぞれ処理結果が MAML 形式で返される．3 章で定義した類似度を基にしたばねモデルを適用して，関連するノードが近傍に配置されるような座標を設定しパネル上に配置する．また，ノード間の類似度を色の濃さで表現している．これによりあるノードに注目した時，そのノードに対する類似度が高いノードほど濃い青色で示し，類似度が低くなるにつれ色も薄くなる．グラフ表示の画面例を図 2 に示す．

4.2.2 クラスタリング処理部

3 章の手法を用いて Web 検索結果のクラスタリングを行う．まず，検索結果として得られた Web 文書のタイトルと抜粋文を一つの文書（アノテーション）とみなし，さらに形態素解析を行い，各文書ごとの単語ベクトルを生

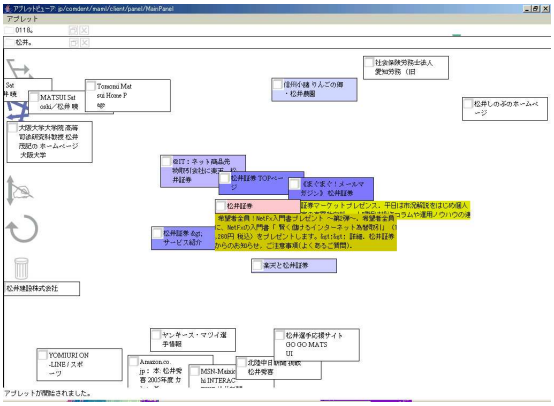

図 2　グラフ表示

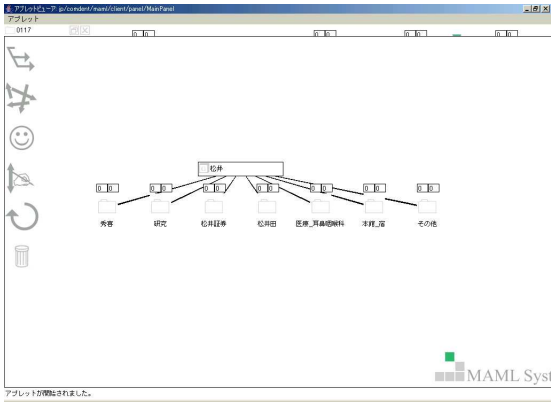

図 4　生成されるクラスタの例

```
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
<maml>
 <media>
  <element id="1"
   annotation="2004/11/03T08:57:14"
   targetURI="u1103549431410-257163406.maml"
   type="homepage">
    <contents>
     <subject>http://www.tokyo-airport-bldg.co.jp/</subject>
    </contents>
  </element>
  <element id="2" annotation="2004/11/03T08:57:14" target="1">
   <contents>
    <explanation>
     羽田空港ターミナル BIG BIRD 公式 WEB サイト
    </explanation>
   </contents>
  </element>
  <element id="3" annotation="2004/11/03T08:57:14" target="1">
   <contents>
    <explanation>
     第 2 ターミナルがオープンしました！ 羽田空港に新しいターミナル
     がオープンし、ますます便利で快適になりました！
    </explanation>
   </contents>
  </element>
       :
 </media>
</maml>
```

図 3　MAML 記述例

表 1　ストップワードリスト

ページ, フレーム, ホームページ, リンク, トップ, 更新, 情報, 検索, サイト, ブラウザ, 表示

成する．形態素解析には形態素解析システム茶筌 [9] を用い，単語ベクトルの要素としては名詞–一般，名詞–固有名詞，名詞–副詞可能，名詞–サ変接続および未知語の品詞を用いた．なお，前処理として文書中の文字表記（全角・半角）の統一，表 1 に示すストップワードの除去を行った．得られた単語ベクトルから文書間類似度を計算し，類似した文書を同じグループとして分類することで Web 文書のクラスタを生成する．その後，結果を MAML System で表示するために MAML 形式のデータに変換する．データの例を図 3 に示す．

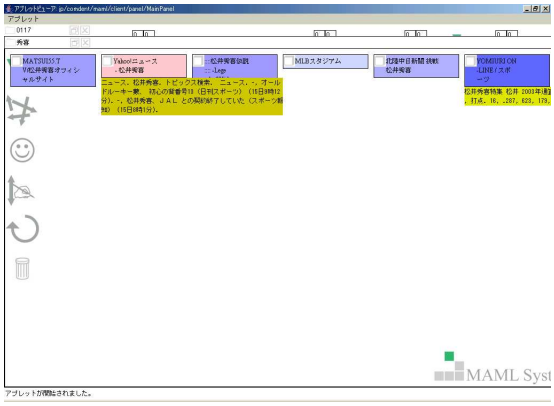

図 5　「秀喜」クラスタに分類された Web 文書群

クラスタリング結果の一例を示す．図 4 では，「松井」という語で検索を行った結果，「秀喜」「研究」「松井証券」「松井田」「医療_耳鼻咽喉科」「本館_宿」「その他」という 7 つのクラスタが生成されている．クラスタとして生成されたフォルダ内には，各 Web 文書の情報（URL，タイトル，抜粋文）がアノテーションとして格納される．図 5 では，「秀喜」クラスタを例として示している．1 つのノードが 1 つの Web 文書に対応している．ノード内のテキストはその Web 文書のタイトルを表している．ノード上にマウスを移動すると，その Web 文書の抜粋文が表示される．

4.2.3　キーワード抽出部

二次元上のある点において，その近傍に存在するノード (以下，近傍ノード) 群の内容を示すキーワードを抽出する．近傍ノードとは，キーワード抽出対象とする点からの距離が th_{dist} より小さいノードをいう．つまり近傍ノード集合を $N_{neighbor}$，キーワード抽出対象の点を n_t，ノード n_i とノード n_j との距離を $dist(n_i, n_j)$，ノード n_i の x 座標，y 座標をそれぞれ x_i，y_i として以下の式を定義する．

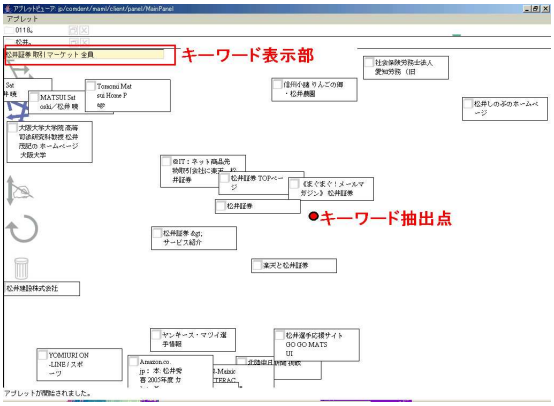

図 6　キーワード抽出例

$$N_{neighbor} = \{n | dist(n, n_t) < th_{dist}\} \quad (8)$$

$$dist(n_i, n_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (9)$$

次に，近傍ノードごとに単語ベクトルを生成する．近傍ノードから，単語の出現頻度とキーワード抽出対象とする点までの距離を取得する．単語 t_k の出現頻度を TF_k，ノード n_i から対象とする点までの距離を d_i とおくと，ノード n_i の単語ベクトル $\vec{v}(n_i)$ を以下のように定義する．

$$\vec{v}(n_i) = (w_1, w_2, ..., w_n) \quad (10)$$

$$w_k = \frac{TF_k}{d_i} \quad (11)$$

次に，各近傍ノードの単語ベクトルの和を求め近傍単語ベクトル $\vec{v}_{neighbor}$ とする．

$$\vec{v}_{neighbor} = \sum_{n \in N_{neighbor}} \vec{v}(n) \quad (12)$$

最後に，近傍単語ベクトルの要素のうち，重みが大きい上位 n_{key} 個の単語をキーワードとして抽出する．結果の一例を図 6 に示す．

4.2.4　フィルタリング処理部

4.2.3 節で述べたキーワード抽出を利用して，検索結果のフィルタリングを行う．検索結果中に含まれるキーワードとすでに配置されている近傍ノードに含まれるキーワードを比較し，検索結果の中から近傍ノードに関連のある Web 文書だけを選択する処理を行う．つまり Web 検索ノードを生成するパネル上の位置によって異なるフィルタリング結果が得られる．

5.　実　　験

5.1　クラスタリングに関する実験

本節では，クラスタリングに関する実験について述べる．

表 2　検索語ごとの再現率・適合率

検索語	再現率	適合率
北朝鮮	0.76	0.83
ダイエー	0.82	0.80
火星	0.90	0.82
羽田空港	0.43	0.54
お花見	0.73	0.79
喜連川	0.77	0.78
松井	0.95	0.98

5.1.1　手　　法

精度を再現率，適合率により評価する．クラスタ C_i についての再現率 $(R(C_i))$，適合率 $(P(C_i))$ を以下のように定義する．

$$R(C_i) = \frac{\text{正しく分類されたホームページ数}}{C_i\text{に分類されるべきホームページ数}} \quad (13)$$

$$P(C_i) = \frac{\text{正しく分類されたホームページ数}}{C_i\text{のホームページ数}} \quad (14)$$

まず，自動的に生成されたクラスタのラベルに対し，人手で該当するホームページを割り当てる．そして，本手法によるクラスタリング結果が一致した場合を正しい分類とする．

5.1.2　結果・考察

実験結果を表 2 に示す．7 つの検索語それぞれ，検索件数は 50 件である．比較的単純なクラスタリング手法にもかかわらず，それなりの精度を示している．しかしながら，検索語によって精度にはばらつきが生じた．例えば「羽田空港」という検索語に対して「空港」というような，あまりクラスタとして意味をなさないものが生成された．それが全体の精度を低下させる要因となった．

5.2　視覚化に関する利用調査

本節では，視覚化に関する利用調査について述べる．

5.2.1　手　　法

提案手法を用いた Web 検索（検索件数 50 件）を行ってもらい，6 項目について 7 段階評価 (1:劣っている~ 7:優れている) をしてもらった．

5.2.2　結果・考察

実験結果を表 3 に示す．本稿の目的である検索結果の閲覧性向上に特に関わるものは「検索結果の概要の把握」と「ページの見つけやすさ」であろう．この 2 項目には，低評価（評価 3）をつけた被験者と高評価（評価 5,6）をつけた被験者が同程度存在した．評価が分かれた理由としては，システムの利便性というものが利用者の好みや慣れに依存する面があるからだと考えられるが，本手法により少なくとも利用者の半数にとって閲覧性が向上しているという点を評価したい．

表 3　各項目の評価値別の人数

項目	評価							平均	分散
	1	2	3	4	5	6	7		
操作性	1	0	0	0	3	2	1	5.0	3.1
見やすさ	0	2	2	2	1	0	0	3.3	1.1
反応の良さ	0	2	0	3	1	1	0	3.9	1.8
ページ配置の精度	0	0	0	0	3	4	0	4.6	0.2
検索結果の概要の把握	0	0	3	1	2	1	0	4.1	1.3
ページの見つけやすさ	0	0	4	0	3	0	0	3.9	1.0

6. ま と め

本稿では，クラスタリングと視覚化を併用した Web 検索結果の提示手法を提案した．これにより検索結果を直感的・大局的に把握した，効率の良い Web 閲覧を行うことが可能となった．今後の展望として，一覧性の改善，ユーザによるノードの再配置の活用などが考えられる．

謝　　辞

本成果は，平成 16 年度 IPA 未踏ソフトウェア創造事業の一部である．IPA（情報処理推進機構）及びプロジェクトマネージャーの名古屋大学　長尾確教授に深く感謝いたします．

付　　録

Google API による検索結果上位 50 件に対して行った，本手法によるクラスタリングの結果を表 4 に示す．

表 4　付　　録

検索語	クラスタラベル	文書数
北朝鮮	拉致	13
	韓国	4
	予選，最終	3
	その他	30
ダイエー	ソフトバンク	9
	機構，再生	7
	再建	5
	紹介	4
	西武	3
	福岡ダイエーホークス	3
	その他	19
火星	接近	9
	探査	5
	惑星	5
	画像	4
	天文	3
	その他	24
羽田空港	駐車	13
	ターミナル	10
	時刻，バス	7
	空港	6
	その他	14
お花見	桜，写真	14
	弁当	6
	スポット	5
	その他	25
喜連川	栃木，温泉	18
	カントリー，倶楽部	4
	栃木	4
	足利	3
	東京大学	3
	合併，氏家	3
	その他	15
松井	秀喜	8
	研究	3
	松井証券	3
	本館，宿	3
	医療，機器	3
	その他	30

文　　献

[1] O. Zamir and O. Etzioni: Web document clustering: A feasibility demonstration, in *Proceedings of SIGIR*, pp.46–54 (1998)

[2] O. Zamir and O. Etzioni: Grouper:A Dynamic Clustering Interface to Web Search Results, in *Proceedings of the 8th WWW Conference* (1998)

[3] K. Kummamuru, R. Lotilikar, S. Roy, K. Singal, and R. Krishnapuram: A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results, in *Proceedings of WWW Conference 2004* (2004)

[4] GoogleBrowser Homepage: `http://www.touchgraph.com/TGGoogleBrowser.html`

[5] Google-enabled visual search Homepage: `http://www.anacubis.com/googledemo/google/`

[6] 伊藤 一成, 斎藤 博昭: 汎用アノテーション記述言語 MAML の提案とその生成・処理プロセス, 情報処理学会論文誌, Vol. 45, No. SIG7(TOD22), pp.137–150 (2004)

[7] 伊藤一成: アノテーションの概念に基づく情報可視化インタフェース, 日本データベース学会論文誌 DBSJ Letters, Vol4, No.1 (2005)(to appear)

[8] Google API ホームページ: `http://www.google.com/apis/`

[9] 茶筌ホームページ: `http://chasen.aist-nara.ac.jp/hiki/ChaSen/`