表記の微細な差異を含む内容が同一である料理レシピの効率的検索

安川 美智子（群馬大学 大学院理工学府）

インターネット上では，僅かに表記の異なる同じ内容の料理レシピが複数のサイトで公開されている．このため，クローリングにより収集した料理レシピの大規模コーパスには表記の微細な差異を含む内容が同一である料理レシピが含まれることがある．現代の食文化を調査し，理解するためには，インターネット上の大量の料理レシピデータを分析する手法が重要となる．本研究では，コーパス中のあるレシピを検索要求として，表記の類似性が極めて高い他のレシピを検索する関連文書検索を効率よく行うことにより，実用的な処理時間で，重複レシピを特定する手法を提案する．また，NTCIR-11 Cooking Recipe Search タスクで構築された英語のテストコレクションに含まれる大規模レシピコーパスを用いた評価実験を行い，提案手法の有効性と効率性を確認する．

Identifying duplicate cooking recipes by using efficient document-by-document search

Michiko Yasukawa (Faculty of Science and Technology, Gunma University)

A cooking recipe corpus constructed by web crawling may contain duplicate or nearly duplicate data. Analysis of duplication, similarity and diversity of cooking recipes is important in the study of food culture. If a piece of recipe data on the Internet was reprinted from a blog entry of the same author of the recipe, these two pieces of data may be included in a recipe corpus and recognized as duplicate or nearly duplicate data depending on the trivial or non-trivial difference between their recipe contents. Manual classification by human assessors would be cumbersome if the size of the corpus is large. In this paper, we propose an efficient recipe-by-recipe search for identifying recipe duplication. For evaluation experiments, we use the NTCIR-11 Cooking Recipe Search test collection.

１．まえがき

料理レシピとは，料理単位に使用食材料とそれらの分量および調理法等を明記した指示書である[1]．本研究では，インターネット上の料理レシピのクローリングにより構築された大規模コーパスを検索対象とし，表記は僅かに異なるが記述されている内容は同一である料理レシピを効率的に検索する手法の提案とその評価を行う．

インターネット上には，僅かに表記の異なる同じ内容の料理レシピが複数のサイトで公開されていることがある．このため，クローリングにより収集した料理レシピの大規模コーパスには表記の微細な差異を含む内容が同一である料理レシピが含まれることがある．表記が完全に同一である 2 つの料理レシピは，ファイル間の差分が無いことを確認する手法により特定できる．しかし，僅かでも差分がある場合には，単純なファイル比較の手法は適用できない．また大規模なコーパスを検索対象とする場合，文章の構造解析等を用いた従来手法は，1 件あたりの計算の負荷が大きく，実用的な時間で全体の検索処理を終えることができないという問題がある．

そこで，本研究では，コーパス中のあるレシピを検索要求として，表記の類似性が極めて高い別のレシピを検索する関連文書検索（文書による文書の検索）を行うことにより，実用的な処理時間で，内容が同一で表記が異なるレシピを効率よく検索する手法を提案する．また，NTCIR-11 Cooking Recipe Search タスクで構築された英語のテストコレクションに含まれる大規模レシピコーパスを用いた評価実験を行い，提案手法の有効性を確認する．

２．関連研究

花井ら[2]は，レシピ検索の妨げとなる酷似レシピの自動抽出を目的として，レシピクラスタリングの手法を提案している．レシピ投稿サイト Cookpad[3]を検索して収集した日本語の料理レシピを用いた評価実験の結果，レシピの材料部分に出現する単語に重みを与えることで，人間の評価者がタイトルを見ただけでは判断できない類似レシピについてもコンピュータによる自動抽出ができたと報告している．

Yasukawa ら[4]は，楽天レシピ[5]に投稿された日本語の料理レシピ約 44 万件を検索対象として，関連文書検索による類似レシピ検索を行っている．また，得られたレシピの類似性を人間の評価者が判定する実験を行っている．

人間の評価者が行う主観的な類似性の判定には，「レシピの表記上の類似性」と「レシピに記載された料理の味や献立における役割（主菜・副菜・デザート等）の類似性」の 2 つの観点があることを Yasukawa ら[4]は報告している．本研究では，「レシピの表記上の類似性」の観点から，人間の評価者が同一性を容易に，かつ，客観的に判定できる範囲の類似性に限定し，表記上の類似性が極めて高い関連レシピの組を，コンピュータを用いて，効率よく自動抽出することを目的としている．英語の料理レシピの大規模コーパスを検索対象として，Yasukawa ら[6]は，レシピ検索サイト Yummly[7]のサーバに蓄積された英語の料理レシピ約 10 万件からなるコーパスを構築している．このコーパスに含まれるレシピの重複や類似性については，これまでに検討されていない．本研究では，このコーパスを後述の評価実験で利用する．

Meng ら[8]は，HTML ファイルの構造解析を行い，ファイル中に含まれる料理レシピの URL を用いて，インターネット上の料理レシピを次々にクローリングする手法を提案している．本研究では，すでにクローリングされて，研究目的で公開されている料理レシピコーパスを検索対象とし，クローリングそのものについては扱わない．

3．予備実験と問題定義

インターネット上で公開されている大量の料理レシピを用いることで，現代の食文化の研究に，文章の計量分析が導入できる．インターネット上の料理レシピには，日常生活で頻繁に調理される料理の食材や調理法が記載されている．食の地域性や季節性，食材の選び方や献立の立て方についての理解を深めるためには，大量の料理レシピデータを客観的に分析し，レシピに記載された内容の同一性，類似性，多様性を明らかにすることが重要である．

料理レシピの内容の同一性，類似性，多様性は，以下のように分類できる．

・レシピの内容が同じ
  [A] 記載が同一（→重複レシピの除去）
  [B] ほぼ同一（→重複レシピの除去）
・レシピの内容が異なる
  [C] 類似性がある（→関連性の分析）
  [D] 類似性がない（→多様性の分析）

表 1　語の出現分布の比較

	RECIPE	CACM
#doc	101,783	3,204
#term	64,864	11,093
size	15,640,350	220,757
#tf(1)	38,093	4,961

（#doc は文書数，#term は異なり語数，size は総語数，#tf(1)は出現頻度 1 の語数を示す．）

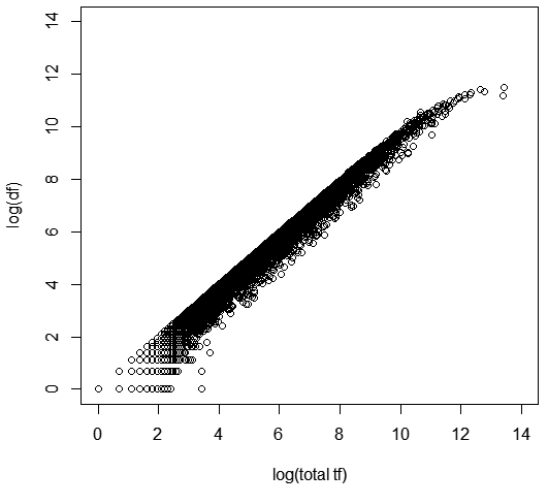

図 1　コーパス RECIPE における語の分布

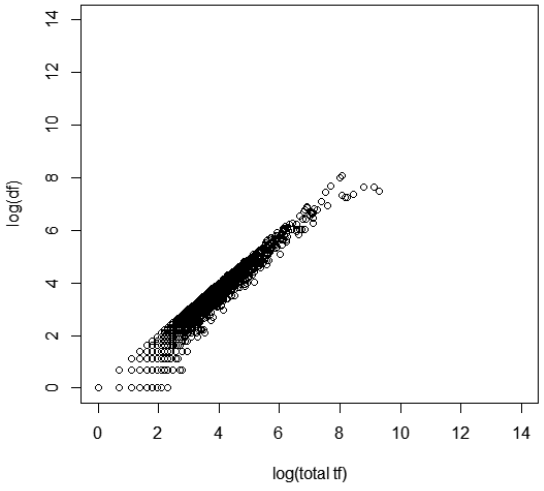

図 2　コーパス CACM における語の分布

本研究で取り組む課題は，大量のレシピデータの中から同じ内容のレシピ（上記の[A]と[B]）を効率よく特定することである．

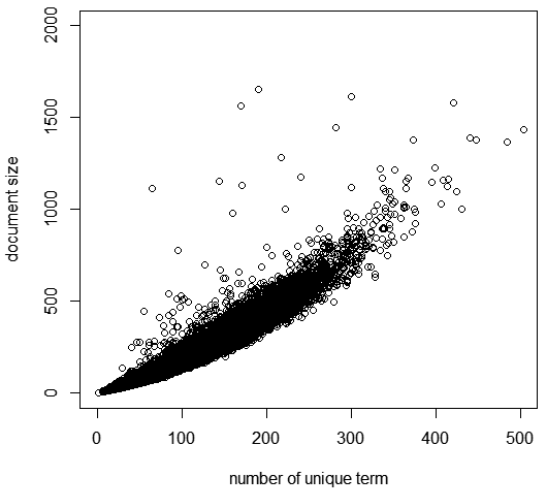

図 3　コーパス RECIPE における文書中の異なり語数と文書サイズ（単語頻度総数）の関係

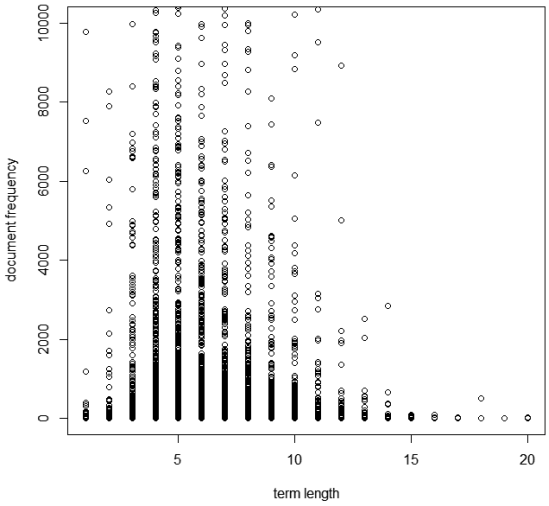

図 5　コーパス RECIPE における語の長さ（文字数）と文書頻度(DF 値)の関係

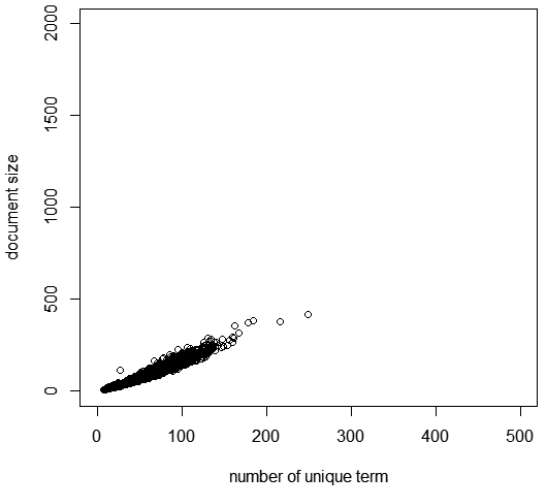

図 4　コーパス CACM における文書中の異なり語数と文書サイズ（単語頻度総数）の関係

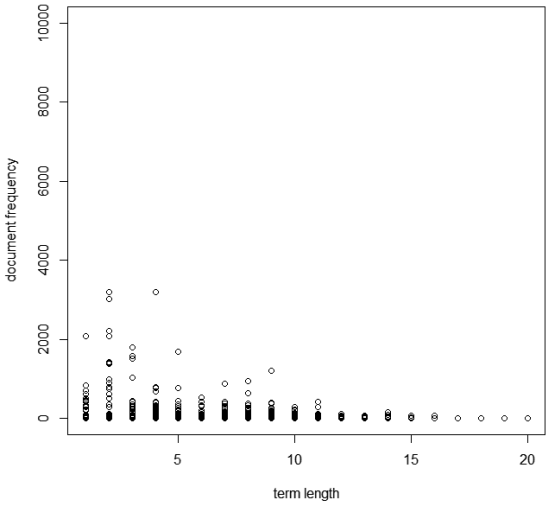

図 6　コーパス CACM における語の長さ（文字数）と文書頻度(DF 値)の関係

具体的には，[A]は，出現する語の順序や回数が同じであり，レシピの記載内容が完全に同一で，表記に差異が含まれないものである．[B]は，出現する語の順序や回数など，表記に微細な差異が含まれるが，人間の評価者が目視で確認すれば，記載された内容が同一であることを，客観的かつ容易に確認できるものである．[C]は料理や献立の関連性，[D]は多様性についての分析であり，これらの分析を行う際にも，[A]と[B]を正確かつ効率よく特定し，レシピの重複を除去することが重要となる．

分析対象のデータ数が少数であれば単純なファイル比較の手法で重複データを特定できるが，コーパス中のデータ数が多くなると，現実的な時間で処理するためには，効率性を考慮した工夫が必要となる．本稿で対象とする課題を数値的に明確にするため，NTCIR-11 Cooking Recipe Search Task の英語の料理レシピコーパス（以下，RECIPE と呼ぶ）と 1958 年から 1979 年の Communications of the ACM に掲載された英語論文の概要から成るコーパス（以下，CACM と呼ぶ）を用いた予備実験を行う．それぞれのコーパスから数字と記号を除去し，英語のアルファベ

ットのみを含む文字列を抽出して，文書中に出現する語とする．コーパス RECIPE と CACM の語の出現分布を表 1，および，図 1～図 4 に示す．コーパス RECIPE の方が CACM よりもコーパスのサイズ，異なり語数が大きい．また，コーパス RECIPE には，CACM よりも文書サイズ(文書中の単語総数)の大きい文書が多数含まれている．コーパス RECIPE と CACM は，ともに語の長さの中央値は 8，最頻値は 7 で，語の長さの平均値は RECIPE が 8.67，CACM が 7.72 である．コーパスに含まれる語の長さ（文字列長）と DF 値(文書頻度；document frequency）の対応関係を図 5 と図 6 に示す．

コーパス CACM から上述の[A]（出現する語の順序や回数が同じである文書の組），および，[B]（出現する語の順序や回数に差異があるが出現する単語が同一である文書の組）を抽出するため，文書 3,204 件のそれぞれについて，文書中の単語をソートして重複を省いた単語リストを作成し，コーパス中の文書の組み合わせ(5,132,808 組)について，単語リストの差分を比較した結果，17 組の重複文書の組み合わせが得られた．ファイル比較の処理に要した処理時間は 489 分 29 秒であった[1]．ファイル比較に要する処理時間は，1 組あたりはごく僅かであるが，コーパスのサイズが大きくなると，文書の組み合わせが膨大な量となり，全体の処理が現実的な処理時間に抑えられなくなる[2]．ファイル比較を行う前に，表記の類似性が無い[D]の集合を除外し，記載に類似性がある[C]の集合をできるだけ少なくして，比較対象となる候補の組を絞り込むことが必要である．

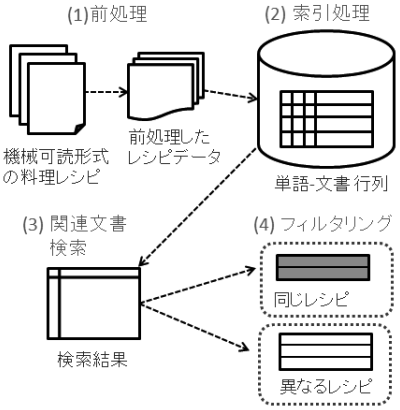

図 7　内容が同一である料理レシピの検索

[1] 処理時間測定に用いた環境の OS は CentOS 6.5 (64bit)，　CPU は Intel ® Xeon ® 2.0GHz, メインメモリの容量は 128GB である．

[2] 1 組あたり平均 0.0057 秒かかるファイル比較を 5,179,889,545 組について行うと 342 日かかる．

4．提案手法

提案手法の概略を図 7 に示す．提案手法は，関連文書検索とフィルタリングを組み合わせることにより，重複レシピの抽出を効率よく行うものであり，以下の 4 つのステップから構成される．

(1) 前処理

HTML や JSON などの記述言語を用いて機械可読な形式で表現された料理レシピから，テキストデータを抽出して，単語に分割し，不要な文字や単語を除去し，索引処理に適した形に変換する．

英語の単語の末尾から不要な文字（接辞）を除去することを目的として Porter Stemmer[9]，Krovetz Stemmer[10]などの接辞処理が提案されている．本研究で提案する料理レシピ検索における接辞処理の有効性を後述の評価実験で確認する．

(2) 索引処理

大規模な特許文書検索などの応用分野で実績のあるオープンソースの検索ライブラリ[11]を利用して，検索効率性が優れた単語-文書行列の構築を行う．

(3) 関連文書検索

コーパスに含まれるレシピ 1 件を検索キーとして，検索キーに含まれる特徴単語を最大 N 個用いて，関連文書検索[12]を行い，検索結果として得られたレシピの上位 M 件を保存する．この関連文書検索をコーパス中に含まれる全てのレシピの数だけ行う．

(4) フィルタリング

検索キーとして用いたレシピとの類似度が 100%である極めて類似度が高いレシピの組のみを比較対象として，単語リストのファイル比較を行い，「同じレシピの組([A][B])」と「異なるレシピの組([C][D])」の分類を行う．

5．評価実験

提案手法の有効性と効率性を確認するために，コーパス RECIPE を用いた評価実験を行った．検索結果数 M=10 とし，特徴単語数 N を 1 から 10 まで step-size=1 で変化させ，さらに 10 から 50 まで step-size=10 で変化させて，関連文書検索を行った．特徴単語の数と類似度が 100%であるレシピの組の数の対応を図 8 に示す．

特徴語の数が少ない場合は，異なるファイル間の厳密な区別がされにくくなり，関連文書検索で得られる高類似度のレシピの組数が多くなる．抽出できた重複レシピの最大組数は，接辞処理なしの場合は 267 組，Porter Stemmer による接辞処理を行った場合は 278 組，Krovetz Stemmer に

よる接辞処理を行った場合は 276 組であり，Porter Stemmer の適用が有効である．

Porter Stemmer を適用した関連文書検索とフィルタリングの結果を図 9 と表 2 に示す．特徴語の数を変化させても，得られる重複レシピ 278 組の組み合わせに変化はなく，特徴語の数の影響を受けるのは異なるレシピの組数の数となっている．表 2 において，特徴語数が 7 個の場合に false negative の誤りが 1 組，すなわち，同じレシピ 1 組が異なるレシピとして誤ってフィルタリングされている．これは，レシピ ID が EN00004631 と EN00071860 のレシピに以下の表記の差異があり，特徴語の出現頻度の差が，文書類似度に影響し，フィルタリングに失敗したためである．

・EN00004631　（材料部分の抜粋）
2 cups brewed or instant tea

・EN00071860　（材料部分の抜粋）
2 cups brewed tea or 2 cups instant tea

接辞処理なしの場合や Krovetz Stemmer による接辞処理を行った場合に false negative となったレシピには以下のような表記の差異があった．Porter Stemmer による接辞処理では，これらの派生形は矢印の右側の表記に正規化される．

- cup/cups → cup
- cook/cooks/cooked/cooking → cook
- curl/curls/curled → curl
- degree/degrees → degree
- egg/eggs → egg
- glass/glasses → glass
- onion/onions → onion
- raisin/raisins → raisin
- serve/serves/serving/servings → serv
- teaspoon/teaspoons → teaspoon

関連文書検索とフィルタリングの処理時間を表 3 と図 10 に示す．表と図は Porter Stemmer を適用した場合を示しているが，接辞処理なしの場合や Krovetz Stemmer による接辞処理を行った場合と比較して処理時間の傾向に大きな違いはなく，特徴語数が多くなると関連文書検索にはより多くの時間がかかるが，フィルタリングの候補となるレシピの組が精度よく絞り込まれるため，フィルタリングの処理時間は短くなる．特徴語数 8 個のときに，有効性，効率性ともに最良の検索を行うことができている．関連文書検索とフィルがリングの処理時間の合計は 11m21.908s

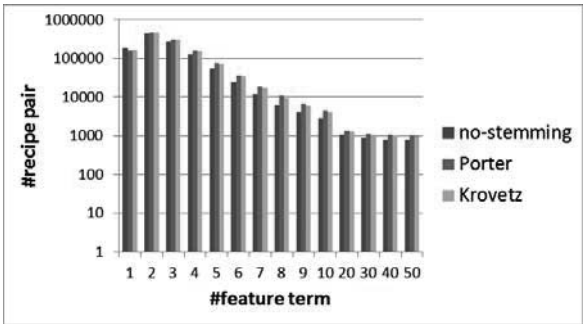

図 8　特徴単語数と検索されたレシピ組数
（横軸は特徴単語の数，縦軸は類似度が 100%であるレシピの組の数，no-stemming は接辞処理なし，Porter は Porter Stemmer による接辞処理を行ったもの，Krovetz は Krovetz Stemmer による接辞処理を行ったものを示している．）

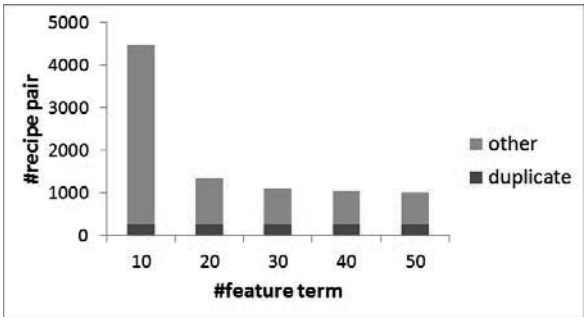

図 9　検索されたレシピの組のフィルタリング
（duplicate は同じレシピの組数，other は異なるレシピの組数を示している．）

表 2　関連文書検索とフィルタリングの結果

特徴語数	関連文書検索	フィルタリング	
		同じ	異なる
5 個	76899 組	278 組	76621 組
6 個	37084 組	278 組	36806 組
7 個	19149 組	277 組	18872 組
8 個	10824 組	278 組	10546 組
9 個	6679 組	278 組	6401 組
10 個	4486 組	278 組	4208 組
20 個	1355 組	278 組	1077 組
30 個	1117 組	278 組	839 組
40 個	1051 組	278 組	773 組
50 個	1004 組	278 組	726 組

であり，予備実験でコーパス CACM に適用した素朴な手法よりも実用的な処理時間が達成できているといえる．

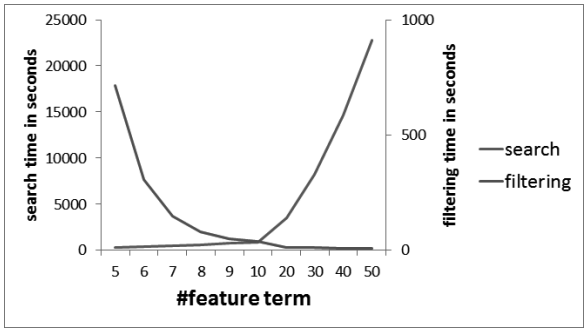

図 10　検索とフィルタリングの処理時間
（search は関連文書検索，filtering はフィルタリングの処理時間を示している.）

表3　関連文書検索とフィルタリングの処理時間

特徴語数	関連文書検索	フィルタリング
5 個	4m24.905s	11m53.188s
6 個	5m56.043s	5m06.456s
7 個	7m52.081s	2m29.205s
8 個	9m50.373s	1m21.535s
9 個	12m15.778s	0m50.644s
10 個	14m48.586s	0m36.129s
20 個	58m07.980s	0m10.980s
30 個	137m32.664s	0m10.170s
40 個	244m01.990s	0m09.252s
50 個	379m33.701s	0m09.081s

６．あとがき

本研究では，表記の異なる同一レシピの検索を行う手法を検討し，英語の大規模コーパスを利用した評価実験を行った. 関連文書検索とフィルタリングの組み合わせにより効率的かつ有効な検索を行えることが確認できた.

本研究では，英語のアルファベットのみから成る文字列を文書中の単語としたが，「2 cups」や「200ml」などの数字や記号で表記された文字列の差異や同一性を文書の類似度計算に含める手法を今後検討する必要がある. また，本論文では，英語のレシピを実験に使用したが，今後の研究で日本語など英語以外の他の言語で書かれた料理レシピについても提案手法を適用し，ストップワード，形態素解析，単語のスペルミスやハイフネーションによる文字列の分割の影響についても検討していく予定である.

謝辞

本研究は JSPS 科研費基盤研究(C)26330363 の助成を受けたものである.

参考文献

1）全国調理師養成施設協会：調理用語辞典（1999）.
2）花井俊介，灘本明代，難波英嗣：スパムレシピ抽出のための酷似レシピクラスタリング手法，情報処理学会研究報告，Vol. 2014-DBS-160，No. 26，pp.1-7（2014）.
3）Cookpad：〈http://cookpad.com/〉.
4）M. Yasukawa, H. Ishii, F. Scholer : Gunma University, Kiryu University, and RMIT University at the NTCIR-11 Cooking Recipe Search Task, In Proceedings of NTCIR-11, pp.508-517（2014）.
5）楽天レシピ：〈http://recipe.rakuten.co.jp/〉.
6）M. Yasukawa, F. Diaz, G. Druck, N. Tsukada. : Overview of the NTCIR-11 Cooking Recipe Search Task, In Proceedings of NTCIR-11，pp.483-496（2014）.
7) Yummly：〈http://www.yummly.com/〉.
8) X. Meng, Y. Li, Q. Li: RecipeCrawler: Collecting Recipe Data from WWW Incrementally, In Proceedings of DEWS2006, 〈http://www.ieice.org/~de/DEWS/DEWS2006/〉.
9) M. Poerter: An algorithm for suffix stripping, Program, Vol. 14(3), pp.130-137, (1980).
10) R. Krovetz: Viewing morphology as an inference process, In Proceedings of SIGIR1993, pp.191-202, (1993).
11) GETA:〈http://geta.ex.nii.ac.jp/〉.
12) GETA: libae チュートリアル，〈http://geta.ex.nii.ac.jp/getaN2001/gdoc/geta/tutorial/libae/section1.html〉.