

オークション検索クリックスルーログからの属性値抽出

水本智也[†] 坂口慶祐[†] 小町守[†] 内海慶[‡] 河野洋志[‡] 前澤敏之[‡] 佐藤敏紀[‡]
[†] 奈良先端科学技術大学院大学 [‡] ヤフー株式会社
[†]{tomoya-m, keisuke-sa, komachi}@is.naist.jp
[‡]{kuchiumi, hkouno, tmaezawa, toshsato}@yahoo-corp.jp

1 はじめに

サイト内を検索することで瞬時に複数の店舗・商品を比較して、安価に商品を購入することができるため、Amazon や楽天市場、Yahoo! オークションなどのオンラインショッピングやウェブオークションが近年広く使われている。しかしながら、出品される商品が膨大なため、複数のトークンを含む検索クエリ^{*1}を用いることで検索結果を絞り込まなければ、比較検討して目当ての商品に到達することが不可能なことが多い。加えて、どのような単語で絞り込めばよいかは自明ではない。たとえば、検索クエリ「iPhone 4S ブラック」を用いると、「iPhone」という商品のうち「4S」という型番の「ブラック」色のものに絞り込んで検索できるが、このような検索は、個々の商品にどのような属性があるかをオークションユーザ自身が知らない限り可能ではない。

このような状況を解消するために、商品が持つ属性を自動的に獲得する、という研究が行なわれている。たとえば [5] は属性名と属性値のペアをシードとする半教師あり学習によって、商品の説明文から属性名と属性値のペアを取得する手法を提案した。この手法を、オークション検索に適用するためには個々の商品に対応した属性値の一覧が必要となる。しかしながら、オークションやショッピングサイトには無数の商品が、非構造データを多く含む説明文とともに出品される。そのため、各商品に対応した属性（属性名と属性値のペア）を説明文から抽出し、それをシードにすることはコストや精度の面で現実的ではない。

一方、実際のオークションユーザの検索ログを分析すると、多数のユーザが「iPhone」などの商品名に加え、「4S」や「ブラック」のような属性値を入れて検索結果を絞り込んでいることが分かる。特に、ユーザがある検索クエリを入力した直後にどのアンカーテキストをクリックしたかという**検索クリックスルー**に関するログデータは、ユーザの検索意図を反映しており、ユーザ経験の向上に有用な知識を獲得できると考えられる。

そこで、本研究は Yahoo! JAPAN のオークション検索

のログから、商品に対する属性値を獲得することを試みた。特にオークションの検索クリックスルーログから、属性名も属性値も与えず、商品名のみからオークション検索のための属性値を抽出する手法を提案する。

2 関連研究

ウェブの検索ログは大規模に入手可能であるという利点を活かし、検索ログからの知識獲得は盛んに研究されている。たとえば検索クエリログからはオントロジー獲得 [6] や集合拡張 [4]、属性名抽出 [3] などに、検索クリックスルーログは意味カテゴリ獲得 [10] スペル訂正知識獲得 [7] に用いられている。また、最近ではセッションログからクラス名を抽出する手法も提案されている [2]。検索の行動ログを用いているという点で我々の研究はこれらの研究と共通しているが、我々の研究はウェブ検索ではなくオークションの検索ログを知識源としており、属性値を抽出するという点で異なる。

一方、属性値抽出に関する手法も盛んに研究されている。たとえば [8] は Wikipedia の Infobox から属性名と属性値のペアを取得する方法を提案した。[9] も [5] と同じく商品の説明文から属性名を用いた統計的検定手法によって属性名と属性値のペアを取得している。属性値を抽出するという点で我々の研究はこれらの研究と共通しているが、我々の研究はオークションの検索ログを知識源としており、収集コストの高い属性名も属性値も用いず、商品名をシードとして与えるという点で異なる。

3 オークションの検索クリックスルーからの属性値抽出

本節では検索クエリとそのクエリが入力されたときにクリックされたアドレスをそれぞれノードとする 2 部グラフを構築し、属性値を抽出する手法について述べる。図 1 に示したように、オークション検索ログからクエリとそのクエリを入力したときにクリックされたオークション ID（クリックスルー）をノードとして、クリックスルーグラフを構築する。エッジの重みはクエリとクリックスルーの関連度を表す。同じオークション ID に到達するクエリは同じ商品を購入する意図で検索された

^{*1} 本研究では空白文字で区切られた各文字列をトークンと呼ぶ。

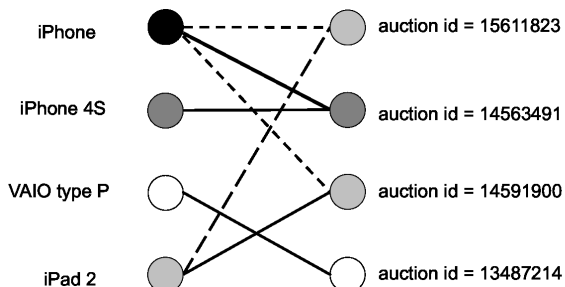


図1 オークション検索クリックスルーグラフ。エッジの線の強さは関連度を表す。ノードの色の濃さは、1トークンクエリ「iPhone」に対する複数トークンクエリの関連度の強さを表す。

可能性が高く、オークションユーザは属性値を用いて絞り込み検索を行なうことで購買活動をしているため、商品名からなる1トークンクエリと、同じオークションIDに到達する2トークンクエリを比較することで、効率的に属性値を列挙することが可能である。

たとえば、図1では、1トークンクエリ「iPhone」で検索したとき閲覧されたオークションIDが3つあるが、そのうちもっとも関連度が高いオークションと関連している2トークンクエリの「iPhone 4S」もまた、「iPhone」で検索して同じ商品を購入したいと考え、「4S」を属性値として獲得する。

本研究では[10]に従い、正規化ラプラシアン $\tilde{L} = I - D^{-1/2}AD^{-1/2}$ を用いた。以下の行列 K とクエリ次元で商品に対応する要素のみ1、他は0となるベクトルの積で商品に対するクエリのスコアを求め、上位となるクエリから順に、商品にマッチするトークンを除いたトークンを属性値として抽出する^{*2}。

$$K = I - \alpha \tilde{L} \quad (0 \leq \alpha < \lambda^{-1}) \quad (1)$$

ただし、クエリ類似度行列 A は $A = W^T W$ とし、クエリ・クリックスルー行列 W の $W_{i,j}$ 要素はクエリ x_i とクリックスルー c_j の正規化自己相互情報量[1]である。 D は A の次数対角行列で、 $D_{ii} = \sum_j A_{ij}$ と定義される。

直観的には、商品名のクエリからオークションIDを介して1ホップで到達できるクエリは、同じ商品に属性値で絞り込み検索をかけたものが含まれる、と仮定するモデルである。しかしながら、属性値は出現頻度が低いいため、単純な頻度を用いると抽出することができない。そこで、正規化自己相互情報量および正規化ラプラシアンを用いることで、頻度が高いクエリの影響を減じることを図っている。

4 オークション検索ログの分析

オークションの検索クエリログは、一般的なウェブ検索のそれとは異なる傾向が見られる。例えば、Yahoo! JAPANにおけるウェブ検索クエリ、オークション検索

表1 Yahoo! JAPANにおけるウェブ検索とオークション検索の2トークンクエリログ比較

1stトークン	クリックコボから		祇園祭	
	ウェブ	オークション	ウェブ	オークション
2ndトークン1	主題歌	試写会	2011	観覧席
2ndトークン2	評価	舞台挨拶	宵山	チケット
2ndトークン3	感想	前売り券	交通規制	観覧
2ndトークン4	あらすじ	原画展	山鉦巡行	ちまき
2ndトークン5	映画館	ポスター	ちまき	手ぬぐい

クエリについてそれぞれ2トークンの検索クエリログを比較すると(表1)、オークション検索クエリログでは1stトークンが商品名もしくはそれに関連する購入可能なものとして扱われ、それに続く2ndトークンはその商品に関する詳細な属性値が出現していることがわかる。このようなオークション検索クエリを詳細にみていくと、「bravia kdl-32cx400」や「ideapad s10-3」のように、半角英数文字列で表現される型番情報などが2ndトークン以降に数多く出現することがわかる。これらの文字列はウェブ検索クエリでは頻度が低いため、抽出するのが難しい。

今回使用したYahoo!オークションでは、25の大きなカテゴリに分けられ、オークション検索クエリログには検索されたクエリだけでなく、どのカテゴリで検索されたかという情報も含まれている。事前に人手で観察した結果、各カテゴリごとのクエリの表層や検索頻度の傾向に違いが見られた。そこで各カテゴリにおけるクエリの傾向を今後の作業で考慮すべきか判断するため、オークション検索クエリログから各カテゴリごとの言語モデルを構築^{*3}してテストセット・パープレキシティ(以後パープレキシティと呼ぶ)を計算した(表2)。なお今回は1週間分(2011/08/08~2011/08/14)のクエリログを使用し、オークション検索クエリの長さをふまえて文字7gram(L-7gm)と単語3gram(W-3gm)を用いた。ベースラインにはカテゴリ情報が付与されていないオークション検索クエリログの言語モデルを準備し、評価には「カテゴリなし」、「コンピュータ」、「不動産」、「ファッション」、および「コミック、アニメグッズ」カテゴリのデータを使用した。各評価セットに対するパープレキシティをみると、文字7gram、単語3gramともに同じカテゴリの言語モデルがもっとも低く、評価セットに近い言語モデルになっていることがわかる。さらに、「家電、AV、カメラ」対「コンピュータ」のように類似しているカテゴリ間では、異なるカテゴリから作られた言語モデルに対してベースラインよりもパープレキシティが低くなる。この結果から、これらの類似カテゴリ間では抽出可能な属性情報も類似していると考えられる。

^{*3} 言語モデルの構築には IRST Language Modeling (IRSTLM) Toolkit 5.60.03 を使用した。 <http://hlt.fbk.eu/en/irstlm>

^{*2} 本研究では、 $\alpha = 0.0001$ を用いた。

表2 カテゴリ別言語モデルにおけるパープレキシティの比較. 訓練データと評価データが同じカテゴリである場合は下線, ベースラインよりもパープレキシティが低くなっているものについてはイタリックで示している.

評価データ	カテゴリなし				コンピュータ		不動産		ファッション		コミック, アニメグッズ	
訓練データ	L-7gm	data size	W-3gm	data size	L-7gm	W-3gm	L-7gm	W-3gm	L-7gm	W-3gm	L-7gm	W-3gm
カテゴリなし	6.8	2.2G	320.3	1.8G	7.3	2811.6	17.1	639.6	5.5	2299.6	9.3	2208.4
カテゴリあり (カテゴリ名)												
アンティーク, コレクション	32.1	13M	3219.5	11M	44.1	6989.2	57.7	1464.8	36.3	3440.0	29.4	<i>1836.1</i>
コミック, アニメグッズ	48.1	4.2M	4254.3	3.5M	62.0	9454.6	294.4	4629.2	48.2	3986.5	<u>6.7</u>	<i>172.4</i>
本, 雑誌	26.9	32M	2514.1	26M	34.1	4644.2	44.4	1710.8	36.7	4036.6	21.4	<i>1161.1</i>
映画, ビデオ	36.5	19M	3623.3	16M	47.5	7739.8	90.3	2668.0	47.9	6621.8	27.5	<i>1790.8</i>
音楽	32.8	28M	3744.5	22M	37.0	6297.1	145.7	3539.8	37.9	5453.0	30.7	2299.3
事務, 店舗用品	46.9	4M	3467.3	3.3M	32.6	4990.7	198.9	3137.6	43.1	4047.0	87.9	5201.1
ファッション	26.3	155M	4548.5	125M	39.3	8215.8	98.9	4702.9	<u>4.8</u>	<i>110.8</i>	47.2	3521.8
アクセサリ, 時計	37.8	29M	4689.5	23M	37.2	7279.8	183.9	4775.4	34.2	<i>2227.3</i>	64.5	4949.4
コンピュータ	37.3	36M	4602.7	28M	<u>5.3</u>	<i>134.0</i>	247.7	4266.3	55.9	6902.5	57.5	5223.6
家電, AV, カメラ	35.0	55M	4357.2	44M	18.7	2203.4	88.5	3534.9	53.3	5410.6	76.2	6796.5
食品, 飲料	63.8	4.3M	4174.7	3.7M	60.7	9300.1	157.4	2869.9	56.8	7272.9	107.6	7396.2
住まい, インテリア	35.3	28M	3623.4	24M	32.1	5328.6	33.7	1089.2	35.1	2896.7	68.4	4983.0
ベビー用品	76.0	2.0M	5105.5	1.7M	71.0	12120.5	763.6	6047.9	44.0	3089.4	147.5	7833.4
ホビー, カルチャー	30.6	33M	3880.8	27M	28.8	4579.8	93.4	3600.5	32.9	3248.6	43.1	3313.9
スポーツ, レジャー	28.1	93M	3813.2	77M	29.4	4762.1	81.8	3289.5	30.7	<i>2087.0</i>	73.1	5979.6
おもちゃ, ゲーム	23.2	87M	3075.0	71M	27.2	4016.0	70.1	2057.6	37.3	3139.1	21.4	<i>956.1</i>
その他	38.3	30M	3617.6	25M	41.2	6416.3	70.7	1935.9	44.6	5004.9	25.6	<i>1385.9</i>
花, 園芸	62.7	4.6M	4301.7	4.0M	56.4	9374.5	101.9	2373.7	64.7	7705.9	134.2	9159.2
自動車, オートバイ	29.0	286M	4448.4	225M	30.2	5238.2	75.1	2964.9	54.1	5252.1	107.0	9984.2
ビューティー, ヘルスケア	48.5	11M	4259.7	9.2M	46.5	8143.9	607.7	5321.6	39.7	3854.8	91.5	6334.0
タレントグッズ	61.5	2.9M	5193.1	2.3M	74.3	11741.2	432.6	5292.7	65.6	7581.8	65.1	4122.2
チケット, 金券, 宿泊予約	47.2	37M	7521.2	29M	64.1	14098.8	55.4	1468.8	85.1	11144.5	48.0	3833.5
ペット, 生き物	72.4	2.1M	4399.4	1.8M	67.6	10008.1	292.0	3318.3	66.1	7531.4	125.3	7754.2
不動産	203.4	70K	4588.6	61K	156.7	13209.4	<u>13.5</u>	<i>73.6</i>	155.3	14339.2	385.6	11277.0
チャリティー	110.2	139K	5081.7	109K	97.4	13494.2	736.5	6174.4	84.6	11051.6	169.7	8168.7

5 クリックスルーログからの型番の抽出実験

5.1 実験データ

本研究の実験では 2011 年 10 月分の Yahoo!オークションの検索クリックスルーログ (頻度 5 以上) および検索クエリログを用いた *4.

実験では, 評価データの用意が簡単な属性値のひとつである型番を抽出対象とした. Yahoo!オークションのカテゴリの中で型番がついている商品があり, 検索数の比較的多いカテゴリである「自動車, オートバイ」, 「コンピュータ」, 「家電, AV, カメラ」の 3 つで評価を行なった. 表 2 に示したように, 「コンピュータ」と「家電, AV, カメラ」の傾向は類似しており, 「自動車, オートバイ」は性質が異なると予想される. 「自動車, オートバイ」カテゴリの中には自動車やオートバイに加えて周辺機器類も含まれているが, 評価データの都合上, 「自動車」のみ使用した. 「コンピュータ」では, PC, タブレット端末や周辺機器があるが, 確実に型番がつけられている「PC」のみ使用した. 「家電, AV, カメラ」カテゴリでは, カテゴリ内でも検索数が多い「テレビ」および「カメラ」を使用した. 以後, それぞれ使用した商品より「自動車」, 「PC」, 「テレビ・カメラ」と呼ぶ. 表 3 に実験に使用したクエリを示す. 各カテゴリから 2011 年 10 月のデータで検索数が多い 1 トークンクエリ上位 10 件を実験に使用した *5.

*4 オークション検索で使われる, aaa か bbb の少なくともひとつを含む「(aaa bbb)」や aaa を含めない「-aaa」といったブーリアン検索表現を含むクエリは, 属性値抽出の結果に悪影響を及ぼす可能性があるため取り除いた.

*5 評価データを取得した際, 商品がなく型番がとれなかったクエリは除外した.

表3 実験に用いたクエリ

自動車	PC	テレビ・カメラ
ジムニー	imac	regza
アルファード	qosmio	アクオス
ヴェルファイア	macbook	wooo
ワゴン R	vaio	ビエラ
プリウス	thinkpad	bravia
ハイエース	pavilion	real
エスティマ	dynabook	lumix
レガシィ	pavilion	サイバーショット
カブチーノ	versapro	exilim
クラウン	toughbook	finepix

評価に使用した型番は Web から取得した. 「自動車」の型式については「クルマ・ポータルサイト Goo-net」のカタログページ *6から抽出した. 「PC」に関しては「インバースネット株式会社」のパソコンスペック情報のページ *7から抽出した. 「テレビ・カメラ」の評価データは「価格.com」の API *8を使用して抽出した.

5.2 ベースライン

ベースラインには検索クエリログの頻度を用いた. テスト対象のクエリを 1 トークン目に含む 2 トークンクエリの頻度順上位 50 件を抽出し候補とした. 例えば, テスト対象のクエリが「vaio」であれば「vaio ノート」, 「vaio VPCEK23」といったクエリを抽出してくる.

5.3 評価と実験結果

実際のオークション検索において, 完全な型番がクエリに使われることは少ないことがわかった. 例えば, 「PC」の SONY の VAIO であれば「VPCEK23FJ/W」という型番があるが, 「VPCEK23」や「VPCE」と一部分のみ入れて商品を検索することが多い. このことから, 今回の実験では前方 4 文字以上一致もしくは, 後方 4 文字

*6 <http://www.goo-net.com/catalog/>

*7 <http://www.inversenet.co.jp/pclist>

*8 <http://apiblog.kakaku.com/KakakuItemSearchV1.0.html>

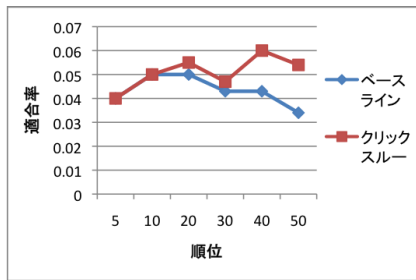


図2 自動車における適合率

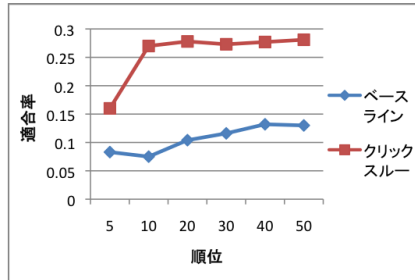


図3 PCにおける適合率

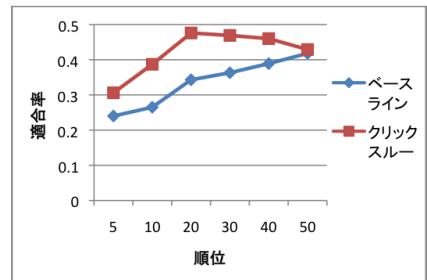


図4 テレビ・カメラにおける適合率

以上一致^{*9}している場合を正解として評価を行なった。

全ての型番を抽出することが困難なため、システム評価には順位 k での適合率 (precision at k) を用いた。適合率は (システムが出力した正解の個数) / (システムが出力したインスタンスの数) である。

図2, 3, 4に各商品において順位 k を変化させた場合の適合率を示す。どの商品を見てもベースラインよりもクリックスルーを用いた手法のほうが適合率が高い。

6 議論とエラー分析

「車」は他の2つに比べて適合率が低くなっている。実験結果を見ると、「タイヤ」、「マフラー」といった自動車のパーツを表すような属性値を抽出している。これはYahoo!オークションのユーザが自動車本体よりも自動車パーツの純正品を探していることが多いことが影響している。オークション検索のユーザパターンは大きく2種類に分けられて、1つは「自動車、オートバイ」といった大カテゴリから「ジムニー タイヤ」と検索するタイプ、もう1つは「タイヤ、ホイール」といった周辺機器の詳細なカテゴリから「ジムニー」といった本体のクエリを入力するタイプである。この2つのパターンのどちらを使っても、同じ商品にたどり着くことができる。「車」に比べると、「PC」、「テレビ・カメラ」は純正品を検索するという傾向が小さいため、このような問題はあまり生じていない^{*10}。この問題は、例えば、本体や各パーツの小カテゴリを利用して別々にグラフを作るなどして、カテゴリ別にグラフを構築することで解決できる。

クリックスルーを用いて抽出した場合、「車」は全てのテストデータで50個抽出することができたが、「PC」、「テレビ・カメラ」はテストデータのうち半分は50個抽出することができなかった。この原因のひとつは、オークションIDをもとにグラフを構築したことであると考えられる。オークションで出品される商品は、時々刻々と変わっているため、クリックスルーのグラフがスパースに

なりやすい。これを解決するひとつの手段として、オークションidではなく、出品タイトルや出品情報を用いて似た商品をまとめることが考えられる。

7 おわりに

本研究では、オークションの検索クリックスルーログを用いて型番情報の抽出を行なった。オークションログを用いることで、Web検索クエリログでは抽出の難しかった属性値を獲得できる事を示した。提案手法は属性名も属性値も必要とせず、商品名のみから属性値を抽出することができ、オークションクリックスルーログを用いる手法の方がオークションクエリログを用いた手法より適合率が高いことを実験で確認した。

「自動車・バイク」カテゴリでは本体より正規品の周辺パーツを検索するユーザが多く、「コンピュータ」カテゴリでは本体を検索するユーザが多いといった、ユーザの商品名を入れた場合の目的の違いから、全ての商品で同じように属性値を抽出することは難しい。また、本研究では型番しか扱わなかったが、他の属性値に適用することが今後の課題である。

参考文献

- [1] G. Bouma, "Normalized (Pointwise) Mutual Information in Collocation Extraction," Proceedings of GSCL, pp.31–40, 2009.
- [2] M. Paşca, "Ranking Class Labels Using Query Sessions," Proceedings of ACL, pp.1200–1209, 2011.
- [3] M. Paşca and B.V. Durme, "What You Seek is What You Get: Extraction of Class Attributes from Query Logs," Proceedings of IJCAI-07, pp.2832–2837, 2007.
- [4] M. Pennacchiotti and P. Pantel, "Entity Extraction via Ensemble Semantics," Proceedings of EMNLP-09, pp.238–247, 2009.
- [5] K. Probst, R. Ghani, M. Krema, A. Fano, and Y. Liu, "Semi-Supervised Learning of Attribute-Value Pairs from Product Descriptions," Proceedings of IJCAI-07, pp.2838–2843, 2007.
- [6] S. Sekine and H. Suzuki, "Acquiring Ontological Knowledge from Query Logs," Proceedings of WWW, pp.1223–1224, 2007.
- [7] X. Sun, D. Micol, J. Gao, and C. Quirk, "Learning Phrase-based Spelling Error Models from Clickthrough Data," Proceedings of ACL, pp.266–274, 2010.
- [8] F. Wu and D.S. Weld, "Autonomously Semantifying Wikipedia," Proceedings of CIKM'07, 2007.
- [9] 宇佐美祐, 萩原正人, 関根聡, "楽天商品データを用いた属性値抽出," 楽天研究開発シンポジウム, 2011.
- [10] 小町守, 牧本慎平, 内海慶, 颯々野学, "ラブラシアンラベル伝播による検索クリックスルーログからの意味カテゴリ獲得," 人工知能学会論文誌, vol.25, no.2, pp.233–242, 2010.

^{*9} 後方一致は後方1文字削除から4文字一致でもOKとした。これは「車」では「DBA-GGH25W」といった型式があるが、最後についている「W」はほとんど意味がなく、ユーザが省いて入力していることが多いためである。

^{*10} regzaでもブルーレイの型番「d-bz500」が抽出されたり、VAIOのDVDドライブの型番「vgp-udrw1」が抽出された。