イベントデータベースとブログの自動対応付け

藤原 泰士† 難波 英嗣† 竹澤 寿幸† 石野 亜耶‡

†広島市立大学大学院 情報科学研究科 〒731-3194 広島県広島市安佐南区大塚東 3-4-1

‡広島経済大学ビジネス情報学科 〒731-0138 広島県広島市安佐南区祇園 5-37-1
E-mail: †{fujiwara, nanba, takezawa, ishino}@ls.info.hiroshima-cu.ac.jp

あらまし インターネットの普及により,ユーザが自由に意見や感想を記載できるブログの数が増加している.その膨大なブログの中から,あるイベントについて言及されたブログを人手で何のイベントなのかを判定し,開催日や開催地などのイベント情報を調べるには多大なコストがかかると考えられる.そこで,我々はまず全国のイベント情報が記載されている Yahoo!ロコからイベント情報をデータベースに登録する.次にブログからイベント名や施設名といった情報を抽出し,抽出した情報をもとにブログをデータベースと対応付けることで特定のイベントに関する詳細な情報を得ることを目的とする.結論として,ブログから抽出したイベント情報を用いてイベントデータベースとブログの対応付けを行う手法を提案し,ベースラインの精度,F 値を向上させた.

キーワード イベント,データベース,ブログ,情報抽出,機械学習

1. はじめに

観光は地域における消費や雇用の増加など幅広い経済効果をもたらすことから,観光立国の実現が目標として掲げられている.また,2007 年 1 月には観光立国推進基本法が施行され,2008 年 10 月に観光庁が設置されるなど,観光を 21 世紀における日本の重要な政策として位置づけた多様な取り組みが推進されている.観光情報の中でも,祭りや展覧会,コンサートなどのイベントに関する情報は,観光客が行動する際の目的となる重要な情報である.観光客がイベントに関する情報を得るための媒体としては,旅行会社などが運営する観光ポータルサイトや,旅行情報雑誌などの観光情報データベースが人手で作成し公開されている.しかし,イベントは全国各地で開催されているため,各イベントに対し,開催場所や開催期間,参加者の評価・感想を人手で収集し,整理するには非常に時間とコストがかかるといった問題点がある.

この問題点を解決するために,観光情報を発信する場としてよく利用されているブログに注目した.ブログにはイベントの評判や感想が記述されているため,イベントに関する情報を得るための有益な情報源である.しかし,ブログにはイベントの評判や感想といった情報は記載されているが,開催場所や開催期間などの詳細な情報は省略されている場合が多い.

そこで本研究では,ブログに開催場所や開催期間などの詳細な情報を付与するために,イベント情報が登録されているイベントデータベースと,ブログを対応付ける手法を提案する.本手法により対応付けられた結果を閲覧することで,ブログからはイベントの評判や感想,対応付けられたイベントデータベースからはブログだけでは省略されているようなイベントの詳細な情報を得ることが可能となる.

本論文の構成は以下のとおりである.2 節では関連研究について,3 節ではイベントデータベースとブログの対応付けについて,4 節では実験,5 節で考察を述べ,6 節で本稿のまとめについて述べる.

2. 関連研究

本節では,本研究に関連する研究を紹介する.2.1 節では,観光支援に関する研究について,2.2 節では,イベント情報に関する研究について,2.3 節では,対応付けに関する研究について説明を行う.

2.1. 観光支援に関する研究

観光を支援する研究として,Ishino ら[1]は広島に焦点を当て,広島の特徴の 1 つである広島電鉄に関する旅行ブログを収集し提示する研究を行っている.

また,藤井ら[2]は,Ishino らの研究をもとに,旅行ブログエントリの観光タイプの自動分類を行い,地図上にマッピングするシステムを構築している.このシステムは,広島 p2walker で公開されている「ぶらり広島電停散歩 MAP[1]」に使用されており,旅行ブログを地図上にマッピングするシステムを構築している.図 1 に「ぶらり広島電停散歩 MAP」の動作例を示す.

図 1:ぶらり広島電停散歩 MAP の動作例[2]

[1] http://p2walker.jp/peace/ja/blog/

藤井らは，ユーザの知りたい情報を効率よく閲覧できるように，旅行ブログを「見る」,「買う」,「泊まる」,「体験する」,「食べる」,「その他」の 6 タイプに分類を行っている．図 1 の右上にあるボタンを選択することで，選択されたボタンに応じた旅行ブログがその内容に対応した点にピンが付与されるようになっている．ユーザはこのピンをクリックすることによって旅行ブログを閲覧できる．

これらの研究は観光情報を扱っているという点で類似しているが，本研究では，イベントに焦点を当て研究を行うという点で異なる．

2.2. イベント情報に関する研究

Nanba ら[3]は，新聞記事と Web からイベント情報の自動抽出を行っている．Nanba らの手法では，新聞記事からイベント名や日時，開催場所などのイベント情報を抽出することで，開催されるイベント情報を収集する．イベント情報を抽出する手法として，人手で収集した手掛かり語を素性に用いて機械学習によるイベント情報の抽出を提案した．本研究でも，Nanba らが提案したイベント情報の抽出手法を用いて，ブログからイベント情報の自動抽出を行う．

吉田ら[4]は，Web からより多くのイベント情報を収集するため，Web ページとブログエントリを用いたイベント情報抽出の手法を提案している．イベント名に対してブログエントリから，イベント名の前後に頻出する表現を抽出する．ブログエントリから抽出された表現を用いてイベント名の収集を行っている．

Sakaki ら[5]は，洪水や自身，台風といった災害情報をイベントと定義し，Twitter を対象に，そのイベントが発生した場所を検出する手法を提案している．Twitter には，ユーザの位置情報が付与されており，その位置情報を用いてイベントの発生地を特定している．

岡本ら[6]は，地名情報を用いることブログエントリからイベント情報の抽出を行う研究を行っている．他のイベント抽出の研究と比べて，そのイベントに対するブログの話題と推移を考慮し，イベント名の抽出を行っている．

イベント情報に特化した関連研究として，島田ら[7]の旅行ブログとイベントデータベースの自動対応付けという研究がある．島田らは，Yahoo!ロコ[2]からイベント名や開催期間，緯度経度といった情報を抽出し，イベントデータベースを作成している．作成したデータベースと旅行ブログを対応付けるために，日付，ブログ内容の類似度，二つの結果の論理積と論理和に基づいて対応付けを行っている．この点に関して本研究とは，イベント情報を抽出し，イベント名の類似度，緯度経度情報を用いた距離を用いるという点で異なる．また，島田らが Yahoo!ロコから作成したイベントデータベースを本研究ではデータセットとして用いる．

2.3. 対応付けに関する研究

相澤ら[8]は，様々な情報集合の中から同一の情報を参照するペアや集合を識別する問題のことを「レコード同定問題」と呼び，解決案の検討を行っている．レコード同定問題には大きく分けて「レコードの重複検出」,「レコードの統合」,「レコード参照先の判定」,「レコード共参照関係の分類」の 4 種類に分類される．本研究ではこの 4 種類のレコード同定問題のうち，テキストデータの参照先を，指定されたデータベース内の登録レコードから発見するという「レコード参照先の判定」に分類される．

レコード同定問題の処理の流れとしては，セグメンテーション・正規化・選別・比較・検証・調整が一般的である．その中でも特徴的な処理として選別と比較が挙げられ，いかにレコード間の照合スコアを求めるかという問題に尽きる．

Christen ら[9]は，バイグラムインデクシング(bigram indexing)と呼ばれる手法を提案している．キーの値を文字単位バイグラムの集合に変換し，ある一定のしきい値以上でバイグラムが一致する単語間のペアを見つけ出すことも目的としている．

Su ら[10]は，動的計画法(DP マッチング)を用いることで，2 つのパターンの要素間の対応付けを行う手法を用いている．対応付けを行う際に重み付けを行い，その値が小さいほど，その 2 つが類似しているペアとして判定することができる．本研究では，DP マッチングの要素として形態素を用いるため，2 つの単語のペアそれぞれに形態素解析を行った結果と使用する．

一定長の N グラムの一致数を用いて類似度を求めることで,類似する候補を求める方法も提案されており，編集距離と整合性がとれた選別法として挙げられる[11]．また，情報検索の分野で広く普及している tf-idf を類似性尺度として単語間の類似度を計算する手法も提案されている[12]．具体的には，tf-idf による距離が一定値以下であるような単語を集めてキャノピーと呼ばれるクラスタを逐次構成することで，類似する単語を導き出すという方法である．

3. イベントデータベースとブログの対応付け

3.1. 対応付けの概要

本研究で提案する，イベントデータベースとブログの対応付けの流れを図 2 に示す．まず，ブログからイベント情報を自動抽出する．次に，全国のイベント情報が登録されたイベントデータベースとブログが対応付くか否かの自動判定を行うことで，ブログに記載されているイベントの詳細情報をイベントデータベースから取得できるという流れとなる．

[2] http://loco.yahoo.co.jp/

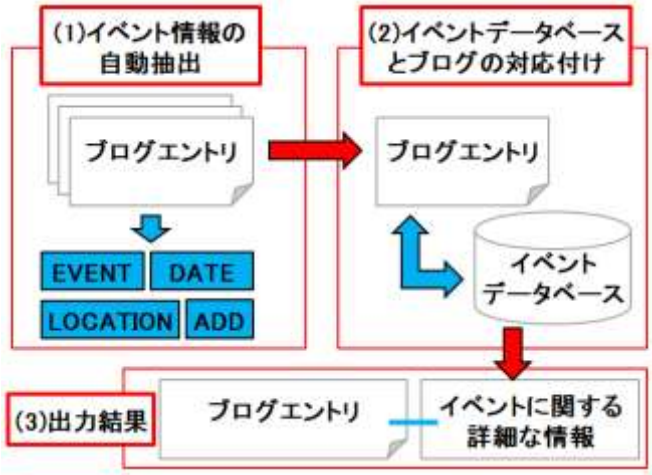

図 2：イベントデータベースとブログの対応付け手順

ブログに対して，イベントデータベースからイベント情報を付与した例を図 3 に示す．このように，イベントデータベースを対応付けることで，ブログには記述されていない，詳細なイベント情報を得ることができる.また,ブログにイベント情報を付与することで，ブログを地図上にマッピングすることも可能である．

<body>
昨日はとうかさんという祭に行きました。人がいっぱい並んでいる出店ではなるべく買いたくないと思ってしまう私…だけど！勇気を出さねばっ！！（｀◇´）お目当ては、一口カステラ☆頑固そうなおやっさんの作るカステラを見て一目惚れ（＊－ω－＊）即買いっ（笑）
</body>
<eventdataDB>
イベント名：とうかさん大祭
開催期間：2013 年 06 月 07 日～9 日
開催施設名：とうかさん圓隆寺
住所：広島県広島市中区三川町 8-12
緯度：34.38920423
経度：132.462584674
</eventdataDB>

図 3：イベントデータベースとブログの対応付けの例

3.2. イベントデータベース

2.2 節で説明した島田ら[7]の研究で構築されたイベントデータベースを用いることで本研究ではイベントの詳細な情報をブログに付与する．イベントデータベースの一例を表 1 に示す．

表 1：イベントデータベースの例

イベント名	2013 広島みなと夢花火大会	第 65 回さっぽろ雪まつり
開催期間	2013 年 7 月 27 日	2014 年 2 月 5 日～11 日
場所	広島港 1 万トンバース	大通公園、つどーむ、すすきの
住所	広島県広島市南区宇品海岸 3 丁目	北海道札幌市中央区
緯度	34.35318	43.05998
経度	132.4703	141.348

3.3. ブログからのイベント情報の自動抽出

本節では，ブログからイベント情報を自動抽出する手法について説明する．2.2 節で述べた Nanba ら[3]の手法を用いて，ブログに含まれているイベント名，住所情報，開催施設名，開催日時といったイベント情報の自動抽出を行う．この手法では，イベント情報に関連する手がかり語を人手で収集し，CRF を用いた機械学習によって，タグの自動付与を行っている．以下にイベント情報であるタグについて示す．

- EVENT タグ

イベント名を示す(例：八王子花火大会，広島みなと夢花火大会，第 59 回さっぽろ雪まつり)

- ADD タグ

イベントの開催地(例：山形県，東京・墨田区，広島県廿日市市宮島口西 1 丁目 5－3)

- LOCATION タグ

イベントの開催施設名(例：鹿児島神宮鹿児島神宮，広島市現代美術館)

- DATE タグ

イベントの開催日時(例：11 月 17 日から 4 日間，6 月 7・8・9 日)

3.4. イベントデータベースとブログの自動対応付け手法

本研究では，3.3 節でブログから抽出したイベント名，住所情報，開催施設名，開催日時を対応付けに利用する．これにより，ブログがイベントデータベースと対応付くか否かの自動判定を行う．以下に，本研究で提案する対応付け手法を示す．

- EVENT 手法

ブログから抽出したイベント名とイベントデータベースに登録されているイベント名の一致度を調べる．具体的には，イベント名に対して，bigram と DP マッチングによる編集距離を用いることで 2 つのイベント名の類似度を計算する．

- ADD 手法

ブログから抽出した住所情報から緯度・経度を計算し，イベントデータベースに登録されている緯度・経度との距離を用いて対応付けを行う．

- LOC 手法

ブログから抽出した開催施設名を用いて，ADD 手法と同様の対応付けを行う．

- DATE 手法

ブログから抽出した開催日時を用いて，EVENT 手法と同様に開催日時を比較し，ブログとイベントデータベースの対応付けを行う．

4. 実験

本研究では，イベントデータベースとブログを対応付ける手法として，以下の 2 段階に分けて実験を行う．

(1) ブログからイベント情報の自動抽出

(2) 抽出したイベント情報を用いた対応付け

4.1 節では，実験に使用したデータの説明，4.2 節では，(1)ブログからイベント情報の自動抽出，4.3 節では，(2)抽出したイベント情報を用いた対応付けの実験とその結果について述べる．

4.1. 実験データと評価方法

本節では，実験に使用するデータの詳細と提案手法の有効性を示すための評価方法を説明する．

・実験データ

実験には，ishino ら[1]の手法により収集した 1,441 件のブログを対象に，ブログで言及されているイベントがイベントデータベースに登録されているイベントなのかを人手で判定した結果を使用した．人手でイベントデータベースと対応付くと判定されたブログが 261 件，イベントデータベースと対応付かないと判定されたブログが 1,180 件となった．イベントデータベースに登録されているイベントと対応付くと判定されたブログの件数を表 2 に示す．

表 2：人手によりイベントに対応付くと判定されたブログの件数

イベント名	件数
ひろしま菓子博	31
宮島水中花火大会	24
宮島かき祭り	19
采女祭	12
とうかさん大祭	8
東大寺二月堂修二会	8
広島みなと夢花火大会	7

・評価方法

評価にはベースラインと 3.4 節で説明した手法を以下のように比較し，評価を行う．

ベースライン

正解データ 1,441 件を全てイベントデータベースに対応付くと判定した結果をベースラインとした．

単体手法

3.4 節で説明した 4 つの手法をそれぞれ単体で用いて，イベントデータベースと対応付くかを判定する．

組み合わせ手法

3.4 節で説明した 4 つの手法をそれぞれ組み合わせることで，イベントデータベースと対応付くか否かを判定する．

評価には，精度，再現率，F 値を用いる．

4.2. イベント情報の自動抽出

3.3 節で説明した手法を用いて抽出したイベント情報の件数を表 3 に示す．

表 3：抽出したイベント情報の内訳

種類	件数
EVENT	116
ADD	289
LOCATION	127
DATE	470

各タグの抽出結果として，合計 1,002 件のイベント情報となるタグをブログに付与することができた．これらを用いて対応付けを行う．

4.3. イベントデータベースとブログの自動対応付け

4.2 節で抽出したイベント情報を用いてブログがイベントデータベースと対応付くかの判定を行う．まず，抽出した 4 種類のイベントタグから 3.4 節で説明した手法を用いて対応付けを行う．その結果を表 4 に示す．

表 4：単体手法の判定結果

	精度	再現率	F 値
ベースライン	0.181	1.000	0.307
EVENT	0.600	0.023	0.044
ADD	0.200	0.192	0.196
LOC	N/A	N/A	N/A
DATE	0.095	0.146	0.115

EVENT 手法で，最も精度が向上するという結果となり，LOC 手法に関してのみ測定不能という結果となった．

次に，最も精度が高かった EVENT 手法に ADD 手法，LOC 手法，DATE 手法を組み合わせて対応付けを行う．その結果を表 5 に示す．

表 5：組み合わせ手法の判定結果

	精度	再現率	F 値
ベースライン	0.180	1.000	0.306
EVENT	0.600	0.023	0.044
EVENT+ADD	0.417	0.345	0.377
EVENT+ADD+LOC	0.404	0.410	0.407
EVENT+ADD+LOC+DATE	0.306	0.421	0.335

結果として，ADD 手法，LOC 手法を組み合わせた EVENT＋ADD＋LOC 手法の場合，精度は若干低下するが，再現率を大幅に向上させ，F 値も一番高くなった．4 つのタグ全てを組み合わせた EVENT＋ADD＋LOC＋DATE 手法は，最も高い再現率だったが，精度と F 値が大幅に低下した．

5. 考察

本節では，4 節で行った実験についての考察を行う．まず，5.1 節で 4.2 節で行ったイベント情報を抽出した結果について，5.2 節では 4.3 節で行ったイベントデータベースとブログの自動対応付けについて考察を行う．

5.1. イベント情報の自動抽出に関する考察

イベント情報である EVENT，ADD，LOCATION，DATE タグそれぞれの抽出結果に対して考察を行う．

- EVENT タグ・LOCATION タグ

EVENT タグと LOCATION タグの抽出結果として，対象としたブログ数に対して抽出したタグ数が EVENT タグ 116 件，LOCATION タグ 127 件という少ない結果となった．本研究では，Nanba らの手法を用いてイベント情報の抽出を行っている．Nanba らの対象とする新聞記事は固い表現が多く，本研究で用いるブログに出現する砕けた表現に対応出来なかったのではないかと考えられる．ブログの砕けた表現に対応した手がかり語を設けることが対策として挙げられる．

- ADD タグ・DATE タグ

ADD タグと DATE タグの抽出結果として，手がかり語として用いた「市」や「区」，「月」や「日」といったものは記事中に多数頻出するため特定のイベントに関連する情報でない場合も抽出してしまう事例が多かった．対策として，ADD タグや DATE タグ周辺に EVENT タグがある場合はそのイベントに対する住所情報や開催期間である傾向にあったため，両者のタグの前後の EVENT タグの有無を素性とすることで改善されるのではないかと考えられる．

5.2. イベントデータベースとブログの自動対応付けについての考察

4.3 節で行ったイベントデータベースとブログの自動対応付けの結果に対して，(1)誤って対応付くと判定された例と，(2)対応付けられなかった例の考察を行う．

(1)誤って対応付くと判定された例

誤ってイベントデータベースと対応付くと判定されたブログの例を図 4 に示す．

<body>
<DATE>７月２８日（土）</DATE>は、全国各地でさまざまな夏祭りがあったようですね。わが家も、この日は、娘のクラブ活動が終わった午後から、夏祭りを、観に、体験しに、出かけました。
**********略**********
その後、夕方からは、当初は広島港の<LOCATION>周辺</LOCATION>で開かれる<EVENT>「広島みなと夢花火大会」</EVENT>を観に行こうと思っていましたが、後日宮島の水中花火大会を観る予定でもあり、今回はあきらめて、地元の商店街で開かれている土曜夜市へ繰り出しました。
</body>

図 4：システムが誤って対応付けした例

内容としては，7 月 28 日に行われた夏祭りの中で広島みなと夢花火大会を見に行く予定であったが，予定を地元のお祭りに変更したという内容である．イベントデータベースに登録されているイベント名と図 4 のイベントタグから抽出される“「広島みなと夢花火大会」”の類似度が高いことや開催期間情報からこのブログをシステムが「広島みなと夢花火大会」に対応付くと判定した．しかし，実際には広島みなと夢花火大会には行かず，イベントデータベースに登録のない地元のイベントに関するブログとなっている．そのため，対応付けを行う際にイベントタグの後に「～が」や「あきらめて」といった手がかり語を用いることで，誤って対応付けを行ってしまうことを防ぐことが出来ると考えられる．

(2)対応付くと判定できなかった例

人手ではイベントデータベースと対応付くと判定されたがシステムでは対応付くと判定されなかったブログの例を図 5 に示す．

<body>
ひろしま菓子博２０１３
今、広島でひろしま菓子博２０１３っていうのが１９日から始まりました♪何となく話の流れで行くことになり出掛けて来ました～。菓子博が始まって初めての日曜なので多いとは思っていたけど・・・
**********略**********
大好きな八ッ橋のチョコ味も見つけたし～～～♪人の多さによってほとんど見る事も無く帰っちゃった～。帰りは駅ビルの麗ちゃんでお好み焼きを食べるつもりがここも並んでて・・・もう並ぶのも嫌だったので隣りのよっちゃんへ・・・・でも結構美味しかったよ！！
</body>

図 5：システムが対応付けできなかった例

システムが対応付けできなかった主な原因としては，ブログからのイベント情報の抽出の失敗が挙げられる．提案手法では，ブログから自動で抽出したイベント情報を利用してイベントデータベースと対応付けを行う．そのため，ブログからのイベント情報の抽出に失敗した場合，対応付けも正しく行うことができない．ブログからのイベント情報の抽出には Nanba らの手法を使用したが，Nanba らは新聞記事を対象として手がかり語を収集し，イベント情報を抽出していた．本研究ではブログを対象としており新聞記事と比べて砕けた表現で文章が記載されている場合が多い．図 5 に関しても「～で開催された」や「～開かれる」ではなく，「～始まりました♪」といったブログ特有の表現が使用されているため，「ひろしま菓子博」というイベント情報を抽出出来なかったと考えられる．そのため，ブログの表現に特化した手がかり語を設けることで，より多くのイベント情報を抽出できるのではないかと考えられる．

6. おわりに

本研究では，ブログに詳細な情報の付与を行うため，イベントデータベースとブログの対応付けを行った．ブログとイベントデータベースを対応付けるために，まず，ブログからイベント情報となるイベント名や住所情報を抽出し，次に抽出された情報を用いて類似度を計算しデータベースとの対応付けを行った．結果として，ベースラインの精度を 0.224 向上させ，F 値に関しても 0.102 向上させることが出来た．

参 考 文 献

[1] Ishino, A., Nanba, H. and Takezawa, T., “Construction of a System for Providing Travel Information along Hiroden Streetcar Lines”, Proc. of the 3rd IIAI International Conference on e-Services and Knowledge Management, 2012.

[2] 藤井一輝，石野亜耶，藤原泰士，前田剛，難波英嗣，竹澤寿幸，“多言語旅行ブログを用いた観光情報提示システム”，第 6 回データ工学と情報マネジメントに関するフォーラム(DEIM 2014)，2014.

[3] Nanba, H., Saito, R., Ishino, A. and Takezawa, T., “Automatic Extraction of Event Information from Newspaper Articles and Web Pages”. Proc. of ICADL 2013, LNCS 8279, pp.171-175, 2013.

[4] 吉田将人，福原知宏，増田英考，“ブログ記事と Web ページを用いたイベント情報抽出手法の提案”，情報処理学会研究報告，デジタルドキュメント 2009(35)，pp.37-44，2009.

[5] Sakaki, T., Okazaki, M. and Matsuo, Y., “Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors”, Proc. of the 18th International World Wide Web Conference (WWW2010), 2010.

[6] 岡本昌之，菊池匡晃，“ブログからの地域イベント情報抽出”，情報処理，Vol.51，No.1，pp.14-17，2010.

[7] 島田恵輔，山本夏生，石野亜耶，難波英嗣，竹澤寿幸，“観光イベントに関する動画とブログの自動収集”，第 6 回データ工学と情報マネジメントに関するフォーラム(DEIM 2014)，2014.

[8] 相澤彰子，大山敬三，高須淳宏，安達淳，“レコード同定問題に関する研究の課題と現状”，電子情報通信学会，Vol.J88-D-I，No.3，pp. 576-589，2005.

[9] Christen, P. and Churches, T., “Febrl - Freely Extensible Biomedical Record Linkage”, Technical Reports, TR-CS-02-05, Australian National University, 2002.

[10] Su, K.-Y., Wu, M.-W. and Chang, J.-S., “A New Quantitative Quality Measure for Machine Translation Systems”, Proc. of the 14th Conference on Computational Linguistics, Vol. 2, pp.433-439, 1992.

[11] Gravano, L., Ipeirotis, P.-G., Koudas, N., Muthukrishnan. S. and Srivastava, D., “Text Joins for Data Cleansing and Integration in an RDBMS”, Proc. of the 19th IEEE International Conference on Data Engineering (ICDE 2003), pp.729-731, 2003.

[12] Cohen, W.-W. and Richman, J., “Learning to Match and Cluster Large High-Dimensional Data Sets for Data Integration”, Proc. of the 8th ACM International Conference on Knowledge Discovery and Data Mining (KDD2002), pp.475-480, 2002.