大規模クロールデータに対する情報抽出の観点からの分析

村上　直也† 　石川　貴大† 　小野　真吾† 　塚本　浩司†

† ヤフー株式会社　〒 107–6211 東京都港区赤坂 9-7-1 ミッドタウン・タワー
E-mail: †{naomurak,takaishi,shiono,kotsukam}@yahoo-corp.jp

あらまし　Freebase や DBpedia など，ウェブ上のデータを元に作られたデータが様々な場面で用いられるようになってきている．これらのデータは人手でメンテナンスされているが，今後構造化されたデータを自動で構築しようとした場合，ウェブ上の情報から機械的に抽出を行なう方法が有望である．しかし，ページが情報抽出元として適しているかは，ページによって大きく異なるため，情報抽出の観点からクロール戦略を立てることが必要となる．そこで我々は，汎用的にクロールした大規模ウェブデータを用いることで，情報抽出に役立つウェブデータにどのような傾向があるか分析を行った．

キーワード　セマンティックウェブ，Web マイニング，情報抽出，Web クローリング

1. はじめに

ナレッジベースと呼ばれるエンティティとその関係を表したデータベースが，ウェブサービスなどに活用されている．例えば “宮崎駿” と検索すると宮崎駿の画像やプロフィールなどが一塊に表示される．この表示内容を得るのにナレッジベースが用いられている．エンティティは**世の中の物事**を表し，関係は**エンティティ間の結びつき**を表す．図 1 にエンティティと関係の例を示す．宮崎駿，スタジオジブリ，もののけ姫がエンティティ，employ，direct，produce が関係となる．また，(宮崎駿，direct，もののけ姫) のような**エンティティ2 つとその間の関係で形成される 3 つ組**をファクトと呼ぶ．情報抽出とは，このようなエンティティ・関係・ファクトを利用可能なデータから抽出することであり，ナレッジベースを構築・拡張するために必要なタスクである．

Freebase [3] や DBpedia [2] などのナレッジベースでは，そのデータの全てまたは一部を Wikipedia の Infobox から抽出している．Infobox が利用されている理由は，データの信頼度の高さと抽出の難易度の低さである．人手で編集された百科事典という Wikipedia の性質上，個々の情報に誤りは含まれているものの全体として見るとデータの信頼度は比較的高い．また，Infobox は半構造化されているので，ルールベースでの抽出が可能であり抽出の難易度は高くないと言える．しかし Wikipedia を情報抽出元とすることには，(1) Wikipedia の成長速度は近年低下傾向にある [8]，(2) 抽出可能なファクトは Wikipedia に存在するものに限られる，という問題がある．そのため，ウェブデータから情報抽出をする必要性は大きい．例えば Knowledge Vault [4] はウェブデータからの情報抽出に特化した抽出システムである．

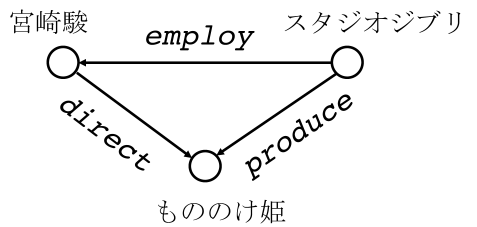

図 1　エンティティと関係の例

ウェブデータからの情報抽出を行うためには，(1) ウェブをクロールしてウェブデータを収集する，(2) ウェブデータの構成要素に対して，各要素に適した手法でファクトを抽出することが必要な作業となる．その際に問題となるのは以下の 3 つである．

- ウェブのクロール方針をどう決めるのか
- ウェブページのどの領域から抽出するか
- どのような手法で抽出するか

このどれもが実際のウェブデータを分析しなければ決められないものである．抽出手法自体は多く発表されているが，我々が調べた限りではそのような分析を行った研究は見つからなかった．

そこで我々は，独自にクロールしたウェブデータについて情報抽出の観点で分析を行った．本研究の貢献は以下となる．

- 大規模ウェブデータを情報抽出の観点から分析
- 情報抽出に役立つウェブページ内の領域を特定
- 情報抽出のためにクロールすべきページの解明

2. 問題の概観

情報抽出に目的を絞らずにウェブをクロールすると，リソースの無駄が多く生じる．図 2 はウェブデータから抽出した文の内，無作為に 10 文を選択したものとなる (注1)．この中で情報抽出に役立つと思われるのは 9 個目の文だけである．この文には “鹿島” と “大迫” というエンティティが含まれていて，“所属選手” という関係を持っている．他の文にはエンティティが含まれてすらいない．このように，情報抽出に役立たないと思われる文が多くウェブデータには存在し，こういった文しか含まないウェブデータを収集することはリソースの無駄である．

(注1)：3. で述べる手法で抽出した文を，記号等も含めそのまま記載している

（1） (10) しょくたく
（2） 予めご了承いただいた上、お申込下さい
（3） 市場環境や資金動向及びその他の要因等により、各ファンドの運用方針が変更されることがございます
（4） ま、こちらは、親切ですね
（5） 防寒性と愛らしさを兼ね備えた 1 点です
（6） 料金が約 1 万円〜なので今度時間が合えば行ってみようかと思います
（7） —のらがし、タイトルについてる
（8） シャツコーデ◆さらりとした肌ざわりでアイロン不要のシャツがおススメ
（9） ビハインドを負った鹿島が積極的に攻撃を仕掛け、後半 2 分に大迫の得点で 1 点差に詰め寄る
（10） 休日・休暇 ※ローテーションで平日週休 2 日制です

図 2　ウェブデータから無作為抽出した文

表 1　ウェブページ上のテーブルの例

映画名	監督	制作
もののけ姫	宮崎駿	スタジオジブリ
白雪姫	デイヴィッド・ハンド	ウォルト・ディズニー

もののけ姫の監督は宮崎駿である．そしてスタジオジブリが制作した．

図 3　ウェブページ上の文の例

よって，役立つページとはどのようなページであるか定義し，その定義に基いてクロール時にウェブページの質の推測ができるようになることが理想である．

実際のデータを分析することで，ウェブデータのどの領域からの抽出を重視すべきかを判断できる．領域とは文やテーブル，箇条書きなどのことである．例えば表 1 のようなテーブルと図 3 のような文では，抽出の手法が抜本的に違ってくる．表では表の持つ構造を抽出に用いることができ，文では表よりも豊かな自然言語の情報を用いることができる．

さらに，各情報抽出手法は入力の種類が決まっているので，たとえ文から抽出する場合でも，文の種類によって抽出手法が限定される．例えば Knowledge Vault のシステムでは，ウェブデータ上の文からの情報抽出に Distant Supervision [6] という手法を用いているが，この手法は 2 エンティティを含む文が必要になる．図 3 の場合，“宮崎駿” と “もののけ姫” の 2 エンティティを含む 1 文目しか用いることはできない．もし 2 エンティティを含む文がウェブデータにあまり存在しないようならば，他の手法を検討する必要がある．しかし分析しない限り，それらについては明らかではない．

3. 提 案 手 法

本研究ではウェブデータが情報抽出に役立つか否かを，**既知のエンティティが含まれるか**という観点で判断する．この観点の利点は情報抽出手法によらないことである．ファクトや関係を用いて分析するためには，特定の情報抽出手法を用いて抽出を行う必要がある．情報抽出手法は多様なため，特定の手法に依存するとその手法の特性が反映されてしまう．一方エンティティの抽出手法は条件付き確率場を使うものが一般的 (注2) であり，特定のツールを選んでも普遍性を保ちやすい．また，ファクトや関係の抽出においてもエンティティを利用するので，エンティティ数だけを考慮してもファクトや関係が含まれているかをある程度反映していると考えた．

ウェブデータの分析は以下の手順で行う．

（1） ウェブデータから，実験項目に合わせてウェブページを選択する

（2） 実験項目に合わせて各ウェブページの領域を抽出する (文，テーブル，箇条書き)

（3） 領域内のエンティティを抽出する

実験項目は 4. で詳しく述べる．テーブルと箇条書きの抽出には HTML のタグを用いるが，文の抽出には次のヒューリスティックを用いる．HTML 中の文字列を分割し，50 文字以内のひらがなを含む文字列を文とする．区切り文字には全角の句点とピリオドを用いる．そしてエンティティ抽出には MeCab を用い，その辞書には IPA 辞書を使用する．MeCab で “固有名詞” とタグ付けされたものをエンティティとして扱う．ただし IPA 辞書にはアルファベットからなる文字列を基本的に固有名詞として登録されていないので，対象となる領域の中に漢字，ひらがな，カタカナのいずれかが少なくとも 1 つのエンティティに含まれている場合のみを実験に使用する．なお，その際に箇条書きは各項を 1 領域として扱う．

4. 実　　験

4.1 分 析 項 目

以下の項目の分析を行う．

（1） ページ単位のエンティティ数

（2） 文単位のエンティティ数

（3） URL の階層の深さとエンティティ数の関係

（4） ウェブページ内の領域の種類とエンティティ数の関係

なお，実験に使用するウェブデータのページ数は約 24 億件であり，再クロールしたページも別のページとして扱っている．再クロールはページに何らかの変更が加わっている場合にのみ行っている．また，(4) 以外の分析では，3. で述べた手法で文が 1 文以上抽出出来たページのみを対象としている．そのようなページは約 2 億 6 千件であり，割合にすると 10.3%であった．

4.2 ページ単位のエンティティ数

ページ単位のエンティティ数を測定する目的は，どれくらいの割合で情報抽出に役立つページがウェブデータに含まれているかを分析することにある．ページ単位のエンティティ数を測定したところ図 4 の結果となり，エンティティを 10 個未満しか含まないページが 84.6%と大半を占めていた．さらに，エンティティを全く含まないページは 40.5%であった．よって，クロールしたページのほとんどはエンティティをほとんど含まな

(注2)：Stanford CoreNLP (http://stanfordnlp.github.io/CoreNLP/) や，本研究で用いる MeCab (http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html?sess=3f6a4f9896295ef2480fa2482de521f6) で採用されている

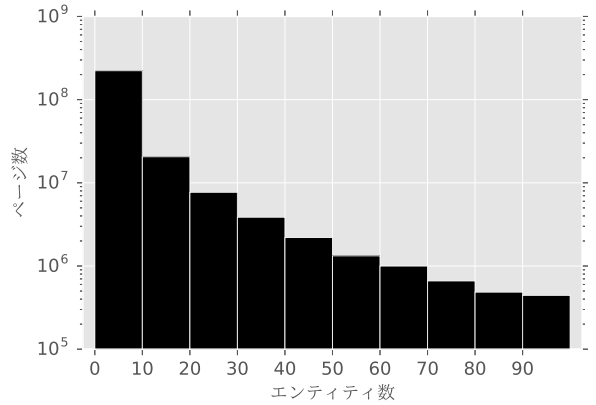

図 4　ページあたりのエンティティ数によるヒストグラム

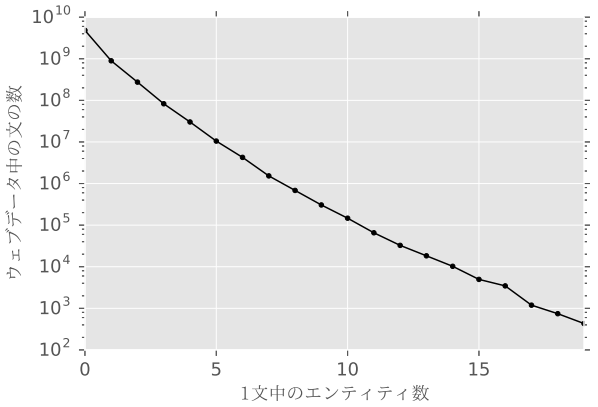

図 5　文に含まれるエンティティ数と文の出現回数の関係

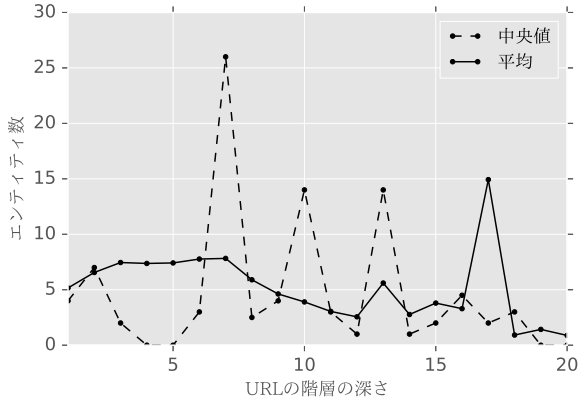

図 6　URL の階層の深さとエンティティ数の関係

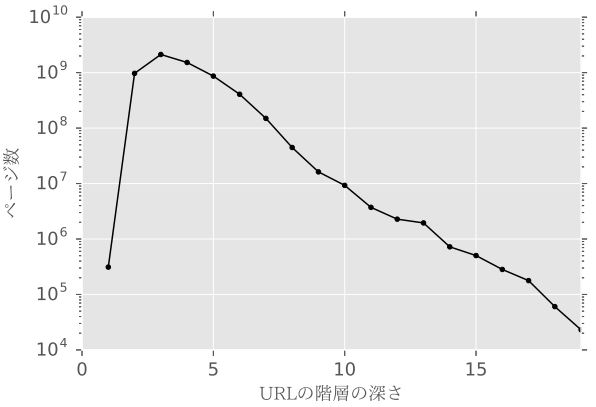

図 7　URL の階層の深さごとのページ数

いページと言える．エンティティを含んだ文の内，情報抽出に有用な文はさらに限られると予想されるため，実際に情報抽出に有用なページはさらに少ないと考えられる．以上より，情報抽出の観点からクロール戦略を立てクロールすべきと結論付けられる．

4.3　文単位のエンティティ数

文単位のエンティティ数を測定するのは，Distant Supervision が適用できるような文，つまり 2. で述べたように 2 エンティティが入っている文がどの程度存在するかの分析を行うことを主目的とする．文単位のエンティティ数を分析したところ図 5 の結果となり，エンティティを 2 個以上含む文の割合は 6.6%であった．また，エンティティを含む文のうち，エンティティを 1 個含む文の割合は 68.8%であった．よって Distant Supervision が適用できないような，エンティティが 1 つしか含まれない文からの情報抽出手法を研究する必要性が大きいと結論付けられる．共参照解析を行えば 1 文中のエンティティを 2 つ以上に増やすことができる文も存在すると考えられるが，そのような文がどれだけ含まれているか分析することは今後の課題である．また，図 2 の 1 つ目の文のように，正確には文と言えないものも入ってしまっていることには留意する必要がある．

4.3.1　URL の階層の深さとエンティティ数の関係

クロールする際に URL のみからエンティティが含まれているかを推測できるのかを知るために，推測手法の 1 候補として URL の階層の深さとエンティティ数の関係を分析する．URL の階層の深さは，**ホスト名以降の部分のスラッシュの数** *+ 1* とする．例えば，`http://db-event.jpn.org/deim2016/` ならば 3 階層とする．このような末尾がスラッシュの場合でも，2 階層目のディレクトリ (この例では `deim2016` ディレクトリ) 内のサーバーで設定されたファイルを読み込むので 3 階層とみなして問題ない．

URL の階層の深さごとの，各ページのエンティティ数の平均と中央値を調べたところ，図 6 のようになった．平均を見ると，あまり階層の深さにエンティティ数は影響されていないが，階層の深さが 17 のときだけエンティティ数が約 15 個と多くなっている．ただし図 7 から分かるように，階層の深さが 3 以降ではページ数が減少し，深さが 17 のときにはページ数が約 18 万件しかない．よってエンティティ数が多い特定のページに平均が大きく影響されている可能性があるので中央値も測定した．深さが 17 のときのエンティティ数の中央値は 2 であり，他の深さに比べてむしろ小さくなっている．また，中央値が大きいときと平均が大きいときが一致していない．以上より，階層の深さとエンティティ数に関係があるとは言えないと結論付けられる．

4.3.2　ウェブページ内の領域の種類とエンティティ数の関係

ウェブページのどの領域から情報抽出すれば良いかを判断す

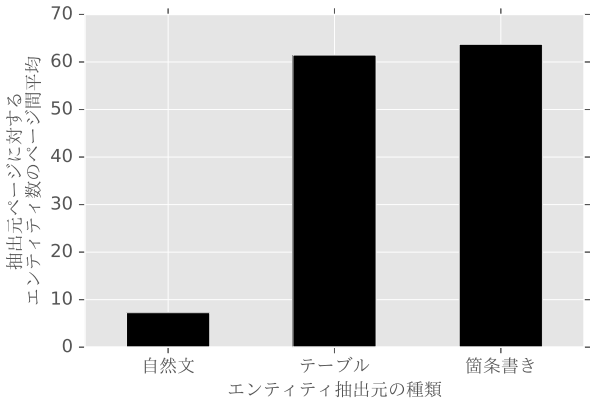

図 8　エンティティ抽出元の種類とエンティティ抽出数の関係
(抽出できたウェブページ間での平均)

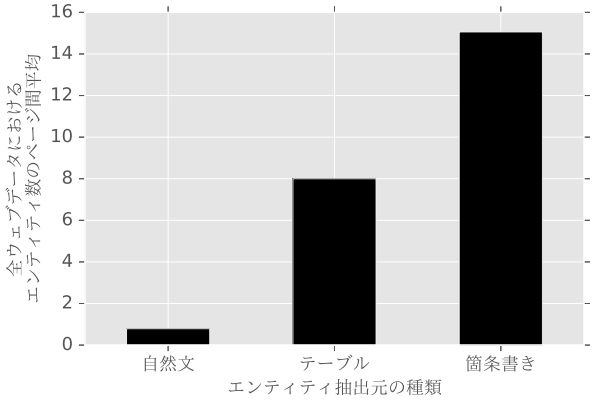

図 9　エンティティ抽出元の種類とエンティティ抽出数の関係
(全ウェブデータにおける平均)

るために，領域とエンティティ数の関係を分析する．分析する領域は (1) 文，(2) テーブル，(3) 箇条書きである．テーブルは<table></table>で囲まれた部分，箇条書きは
で囲まれた部分のこととする．その領域内に他の種類の領域が含まれていても，それも含めてエンティティ数を計測する．例えばテーブル内に文が含まれていた場合，その文中のエンティティも含めて計測する．

計測した結果は図 8 と図 9 のようになり，テーブルと箇条書きから多くエンティティが取れた．全ページ数を p_{all}，種類 k の領域を含むページ数を p_k，全ページ中の種類 k の領域内エンティティ数を e_k とすると，図 8 の縦軸は $\frac{e_k}{p_k}$，図 9 の縦軸は $\frac{e_k}{p_{all}}$ となる．図 8 に比べ，図 9 で箇条書きとテーブルの差が大きいのは，テーブルを含むページ数が少ないからである．箇条書きのエンティティ数が大きくなっているが，箇条書きがウェブページのヘッダーやフッターに使われていることもある．よって箇条書きが情報抽出元として優位だとは言いきれないが，箇条書きが軽視できないことは確かである．以上より，現状では自然文からの情報抽出に関する研究が主流だが，テーブルと箇条書きからのエンティティ抽出の重要性も無視できないと言える．

5. 議　　論

4. で得た情報抽出の観点から重要な知見は次の 3 つである．

- エンティティが抽出できるページは限られている
- 箇条書きやテーブルにエンティティが多く含まれている

箇条書きやテーブルから多くのエンティティが得られることが明らかになったため，クローラーの視点から言えばそれらが多く含まれるページをクロールすることが重要である．上に述べたような知見が得られた一方で，本研究の手法にはいくつかの問題点が見られた．まず，エンティティの抽出手法に改善の余地がある．MeCab で固有名詞として判定される語は，辞書に固有名詞として登録されている語と未知語の一部である．未知語を固有名詞として判別するかは定義されたルールによってなされる．本研究で用いた IPA 辞書では固有名詞として含まれている語が少ないため，偽陽性も偽陰性も多く発生している．偽陽性の例を挙げると，アルファベットからなる文字列は辞書に含まれていないため固有名詞と分類されるケースが存在する．偽陰性の例を挙げると．“グーグル” は辞書に含まれていないため一般名詞と判断されるケースが存在する．よって mecab-ipadic-NEologd[注3] のような，よりエンティティを正確に抽出できる辞書を用いるべきである．ただし本研究のような傾向を分析する目的であれば大きな影響はないと考えられる．

また，エンティティが含まれるか否かでウェブデータの有用性を判断することに起因する問題も明らかになった．4.3.2 で述べたように，ウェブページのヘッダーやフッターのエンティティも用いてしまっている．ヘッダーやフッターからエンティティを抽出できても，エンティティ間の関係を推測・抽出できなければ情報抽出タスクには有用とは言えない．よって情報抽出の観点からウェブデータのより正確な分析をエンティティを用いて行うためには，ウェブデータのクレンジングを十分に行う必要があると言える．

6. 関 連 研 究

本研究のような情報抽出に関する分析を行った研究は我々の知る限りでは存在しないので，類似している研究を紹介する．

Barrio らは，ディープウェブのコレクションを情報抽出の観点でランク付けし情報抽出を効率的に行う手法を提案している [1]．コレクションとは特定のサイトごとに得られるディープウェブのデータのことであり，あるレビューサイトから得たデータが 1 コレクションとなる．コレクションの一部に対して情報抽出手法を適用し，コレクション全体の有用度を推測することで効率的にランク付けを行っている．コレクションの一部の選択には，サンプリングまたはコレクション中のページを取得するクエリを限定することで行っている．情報抽出に適したページは少ないため，ランダムにサンプリングまたはクエリを選ぶと全体の推測に十分なデータが取れない恐れがある．その対策として，情報抽出に向いているページが取れるようにサンプリングやクエリの選択にバイアスを掛け，そのバイアスを考

(注3)：https://github.com/neologd/mecab-ipadic-neologd

慮して全体の推測も行っている．情報抽出の観点からデータを分析している点は本研究と類似性があるが，クロールし終えていることが前提となっている点と情報抽出の効率化が目的な点が異なる．

Dong らはウェブデータの以下の要素に対して，それぞれに適した手法で情報抽出を行うシステムを提案している [4]．

- 文
- テーブル
- DOM
- アノテーション (Schema.org(注4))

そのシステムによる抽出結果として，各要素からの抽出数の違いを明らかにしている．実際に抽出まで行っているので，我々とは異なりファクト単位で分析できている．

Dong らは情報抽出のソースとなるウェブデータの信頼性を推定する手法を提案している [5]．多くのウェブデータに対して複数の抽出手法で抽出した結果を用いることで，ウェブデータの信頼性を求めている．その際に，情報抽出によるエラーとウェブデータ自体の間違いを分離することに成功している．本研究では MeCab のみを用いたが，複数のエンティティ抽出手法を用いれば本研究でもこの手法を用いることが可能である．

7. おわりに

ウェブデータからの情報抽出の必要性が高まっているが，そのウェブデータを情報抽出の観点から分析した研究は見当たらない．そこで本研究では汎用的にクロールして収集した大規模ウェブデータをエンティティに基いて分析を行った．それによって情報抽出手法の選択やクロールの方向性の決定に役立つ分析結果が得られた．より詳細な分析をして具体的なクロール戦略を立てることと，分析に留まらず大規模日本語ウェブコーパスも作成することが今後の研究課題である．

文　献

[1] Pablo Barrio, Luis Gravano, and Chris Develder. Ranking deep web text collections for scalable information extraction. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 153–162. ACM, 2015.

[2] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web*, Vol. 7, No. 3, pp. 154–165, 2009.

[3] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247–1250. ACM, 2008.

[4] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge Vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 601–610. ACM, 2014.

[5] Xin Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. Vol. 8, pp. 938–949. VLDB Endowment, 2015.

[6] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 1003–1011. Association for Computational Linguistics, 2009.

[7] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[8] Bongwon Suh, Gregorio Convertino, Ed H Chi, and Peter Pirolli. The singularity is not near: slowing growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, p. 8. ACM, 2009.

(注4)：https://schema.org/