4B1-R-2-3

能動学習を用いた変遷情報の意味論的統合

Active Learning for Temporal Entity Matching

森田 大翼[*1] Daisuke Morita　飯塚 京士[*1] Kyoji Iiduka　村山 隆彦[*1] Takahiko Murayama　赤埴 淳一[*2] Jun-ichi Akahani

[*1] 日本電信電話株式会社 ソフトウェアイノベーションセンタ
Software Innovation Center, NTT Corporation

[*2] NTT アドバンステクノロジ株式会社
NTT Advanced Technology Corporation

Many datasets contain *temporal data*, which describes some aspects of real-world entities at a certain time. However, histories of data referring to a specific entity are often not stored in such datasets. As a result, temporal data remain underutilized as knowledge from their analysis and statistics. To identify data that describe the same entity over time, we developed a method based on supervised learning, where a classifier was constructed from feature vectors considering changeability of relationships of temporal data with those surrounding entities. Additionally, to improve the practicality by eliminating manual labor of collecting training data, we apply a technique of active learning to temporal entity matching problem. Specifically, considering the feature of temporal data, we propose to add the concept of time distance to existing criteria of selecting queries for labeling unlabeled data. The effectiveness of our approach is confirmed experimentally.

1. はじめに

近年, 組織内に蓄積されている電子情報は増加の傾向にある. その蓄積されたデータの検索や分析を通して, 情報をナレッジとして再利用するニーズが高まっている.

データをナレッジとして再利用するためには, 様々なデータソースに含まれるデータ群を, 実世界での同一性に基づいて統合する必要がある. この技術はデータベースの研究分野では「データ統合」(英語では data integration, record linkage, entity matching などと表記される) 技術として大きなトピックである. Elmagarmid らのサーベイで近年のデータ統合研究が良くまとめられている[Elmagarmid et al. 2007].

実際には, 多くのデータセットは時間情報を伴うデータを格納している. そのデータは, 実世界の実体のある時点に関する情報を持っている. しかし, 多くのデータセットではそのデータの継承関係が管理されていない. 従って, そのデータの履歴を活用するためには, ある期間をまたぐデータを同義性の関係に基づいて統合する必要がある.

例えば, DBLP という数十年に渡る研究論文の情報を蓄積しているデータベースが存在する. しかし, 同姓同名を持つ著者が区別されていないため, 個別の著者に対して論文を列挙できない問題がある. 表 1 に DBLP における Hiroyuki Sato の氏名を持つデータの一部を示す. 表 1 のデータは, 異なる実世界の 3 人の人物$\{r_1 - r_4\}, \{r_5 - r_6\}, \{r_7\}$に分類される. $\{r_1 - r_4\}$は Shinshu Univ から Univ of Electro-Comm.に移動した人である. $\{r_5 - r_6\}$と$\{r_7\}$は所属が同じである, 別の人物である. このような時間情報を持つデータが統合されることにより, 長期に渡る変化を捉えた有益な情報分析を可能にする[Weikum et al. 2011].

従来研究の多くは, データの属性値の文字列類似度を根拠とした手法が主であった. しかし, 時間情報を伴うデータの場合, 表 1 のr_1, r_2のように所属組織などの属性情報が全く異なる場合がある. この場合, 属性間の文字列類似度は非常に小さく, 同義判定を困難にさせる. 我々は, 時間を超えてデータが同義であるような関係性を「変遷を伴う同義性」と定義し, 上記の困難さを克服する判定手法を開発してきた[森田他 2011].

表 1:DBLP のデータ(一部の共著者は削除)

ID	Name	Affiliation	Coauthors	Year
r_1	Hiroyuki Sato	Shinshu Univ.	Aguirre, Tanaka	2009
r_2	Hiroyuki Sato	Univ. of Electro-Comm.	Aguirre, Tanaka	2010
r_3	Hiroyuki Sato	Univ. of Electro-Comm.	Hattori, Takadama	2010
r_4	Hiroyuki Sato	Univ. of Electro-Comm.	Takadama, Otaki	2011
r_5	Hiroyuki Sato	Mitsubishi Electric Corp.	Aoyama, Nakajima	1999
r_6	Hiroyuki Sato	Mitsubishi Electric Corp.	Izumi, Nakajima	2002
r_7	Hiroyuki Sato	Mitsubishi Electric Corp.	Yokotani, Nakatsuka	1993

本判定手法は教師あり学習に基づく手法であり, 教師データ作成のための人的コストが発生する問題がある. 本稿では, この人的コストの削減のため, 能動学習(Active Learning)という少ない教師データで精度を高めることを目的とする技術の変遷を伴う同義性判定問題への適用を試みる. 本稿では, 時間情報を持つデータの性質を考慮した, 能動学習技術の改良を行う.

本稿は以下の通りに構成される. 2 章では, 変遷を伴う同義性判定問題の定義と, アプローチ及び解決手段の概要を示す. 3 章では, 能動学習の概要を示し, 本問題に対する能動学習技術の改良について説明する. 4 章では改良後の能動学習手法の精度の向上度合いについて実験的に明らかにする. 5 章では本実験を受けた考察を述べ, 6 章でまとめる.

2. 変遷を伴う同義性判定問題

変遷を伴う同義性判定問題は, 時間情報を伴うデータをマッチングし, 同じクラスタにあるデータは実世界で時間を超えて同じ実体であるよう, 異なるクラスタにあるデータは実世界で異な

連絡先：〒180-8585 東京都武蔵野市緑町 3-9-11, morita.daisuke@lab.ntt.co.jp

る実体であるようにすることである. 以下, 本問題に対する既存のアプローチと本研究のアプローチを示す.

2.1 先行研究

Liらの temporal record linkage [Li et al. 2011]は調べた限りで見つかっている本問題に取り組む唯一の先行研究である. この研究では, 時間経過に伴い属性値が変化していく傾向を時間減衰(time decay)のメタファで捉え, それに基づいた文字列比較の類似度の調整を行なっている.

しかし, 属性値は変更があるとき, 全く異なる文字列に変更する場合がほとんどである. 例え時間減衰の概念を導入したとしても文字列類似度を判定基準として適用することは適切ではない.

2.2 課題とアプローチ

2.1 節で示した先行研究における問題から, 変遷を伴う同義性の判定問題における課題は以下の通りである.

課題: 属性値に変更があるとき, 全く異なる文字列に変更される場合が多いデータ対に対して同義性を判定すること

本研究では, この課題に対するアプローチのため, まず以下に示す時間情報を伴うデータに対する特徴を仮定する.

仮説: 実世界における実体及び実体間の関係性は穏やかに変化する

例えば表 1 のr_1とr_2は, 所属組織として関わってきた組織の実体との関係は変化したが, 共著者として関わる人の実体との関係は 1 年では変化していない. 仮説の元での変遷を伴う同義性判定問題に対するアプローチは, 1) 2 つのデータに関わる共通の実体を探索し, 2) それらの実体との時間経過に伴う関係性の変化または不変化の特徴を用いることである.

2.3 解決手法の概要

(1) 前処理としてのデータソースの連結

まず, 実体間の関係性抽出のための前処理として, 様々なデータソースを連結し, 一種のデータの Web を作成する. 例えば DBLP は著者データ以外に学会データも保持している. 各論文にはキーワードや, 著者の所属組織情報が含まれている. また, 近年急速に拡大している Linked Open Data [Bizer et al. 2009]という公開データに連結して利用することもできる.

(2) 意味論に基づく特徴ベクトル抽出

本研究における意味論とは,「2 つのデータ間の関係」として定義する. 具体的には, r_1と Shinshu Univ.の間の意味論は「所属組織」である. データの Web の探索により, ある 2 つのデータに共通に関わるデータを取得できる(アプローチ 1). 例えば, 表 1 の Aguirre という人物(Shinshu Univ.に所属)は, r_1とr_2に共通に関わる人のデータである. 次に, その共通に関わるデータに対する意味論の変化を調べる(アプローチ 2). 例えば, r_1の Aguirre との意味論は「共通の所属組織である同僚で且つ共著者」であるが, r_2にとっては「共著者」という意味論であるように, 1 年で意味論が変化している. このような特徴を捉え, ベクトル値化する. 詳細の手順は[森田他 2011]を参照されたい.

(3) 判定モデルの機械学習と適用

一部の正例及び負例となるデータ対の特徴ベクトルを用いて, SVM などの教師あり学習を行い, 同義性を持つデータ対の意味論の変化及び不変化の傾向を学習する. しかし, 仮説より, 同義性を持つデータ対は, その期間差が大きくなるにつれて, 共通性は小さくなる. 従って, 判定モデルの学習に用いるデータ対及び, モデルを適用するデータ対は, その時間間隔が出来る限り小さいことが良い戦略となる. 例えば, 表 1 のr_1とr_4は共通性に乏しいが, r_1とr_2, r_2とr_3, r_3とr_4それぞれのう同義性判定により, r_1とr_4を同一クラスタに分類することができる.

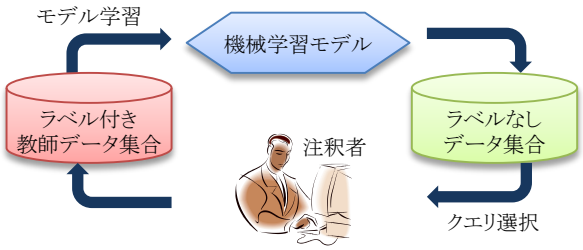


図 1:能動学習のサイクル

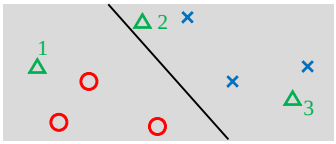


図 2:SVM における不確実さ(○は正例, ×は負例となるラベル付き教師データ. △はラベルなしデータ集合)

3. 能動学習を用いた変遷を伴う同義性判定

2 章で概説した解決手法では, 教師データを用意する人的コストがかかる. 人的コストを抑え, 本技術をより実用的にするため, 少ない教師データで精度の高い判定を行うための技術である能動学習の, 変遷を伴う同義性判定問題への適用を検討する.

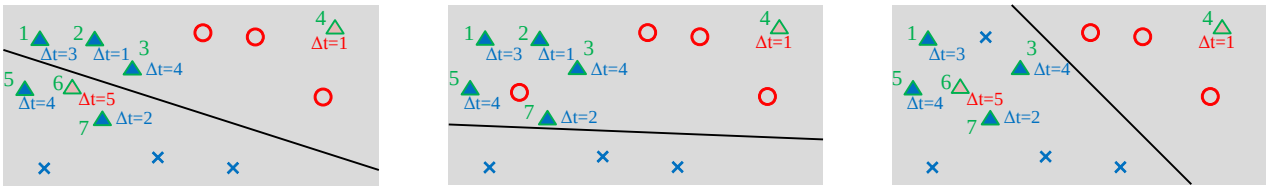
3.1 能動学習

図 1 に能動学習のサイクルを示す. 能動学習は, ある問題に対して事前にクラス分類された, 少量のラベル付き教師データ集合から最初の学習モデルを生成することから始まる. システムは, そのモデルを改善するために最も重要だと推測した数個のデータを人間にクラス分類させるクエリとして選択し, 注釈者にそのデータに正例か負例かのラベルを付けさせる. そして, 追加の教師データを含む集合から新たなモデルを作成する. 性能の高いモデルが生成されるまで, このサイクルは繰り返される.

能動学習研究は, 図 1 のプロセスの「クエリ選択」の良いアルゴリズムを検討して注釈者の負担を可能な限り小さくすることを目的とする. 良いクエリ選択とは, 機械学習のモデルに最も情報量の多いデータを選択することである. その基準として広く用いられる尺度は, モデルに対する不確実さである. 図 2 に SVM における不確実さの直感的な例を与える. 図 2 のラベルなしデータ集合の 1 と 3 に比べ, 2 は SVM の決定境界に対して曖昧である. 既存研究の多くは, この不確実さの尺度をクエリ選択の基準の一つとして用いている[Mirroshandel et al. 2011, Schohn and Cohn 2000]. なお, 他のクエリ選択基準を含む能動学習の既存研究は Settles のサーベイでまとめられている[Settles 2010].

3.2 課題とアプローチ

しかし, 変遷を伴う同義性判定問題に対しては, 時間情報に基づく特徴から, 不確実さの指標だけではクエリの選択基準として適切ではない. 理由は, 2.3 節(3)で述べた通り, 仮説より同義性を持つデータ対であってもその時間差が大きくなるにつれて共通性が小さくなり, 同義性を持たないデータ対と特徴ベクトルが類似する傾向があるためである. 図 3 にこの問題の例を示す. 図 3(a)の状態において, ラベル付けされていないデータ対 1〜7(4 と 6 は実際には同義性の関係となる)からクエリを選択するとする. 不確実性のみを考慮した場合, 超平面に最も近い 6 が選択されラベル付けされる. モデルが再学習された結果は図 3(b)



(a) ある超平面が作成された状態　(b) 時間情報を考慮しないクエリ選択　(c) 時間情報を考慮したクエリ選択

図 3:時間間隔を考慮したクエリ選択(Δt は，データ対の時間間隔を示す)

のようになり，即ち，時間間隔の大きい正例の特徴ベクトルが負例に近いために，モデル学習のノイズとなって汎化性能を低下させてしまっている．以上より，変遷を伴う同義性判定問題の能動学習における課題は以下の通りである．

課題:時間間隔の大きさが原因となり，汎化性能の向上にノイズとなるデータ対をクエリとして選択することを避けること

本課題に対するアプローチは，時間間隔の小ささをクエリ選択の基準の一部として採用することである．このアプローチにより，例えば図 3(a)の状態から不確定度と時間間隔の両方を考慮して時間間隔が 1 であるデータ対 2 を優先的に選択させることで，図 3(c)のように汎化性能を改善できると考えられる．

3.3 解決手法

アプローチを実現する手法として，時間間隔を反映した特徴ベクトルを作成する手法と，時間間隔をクエリ選択の尺度の一部として採用する手法の 2 つがある．以下，各手法の概要を述べる．本研究では実用性の観点から後者の手法を採用する．

(1) 時間間隔を特徴ベクトルに反映する方法

手法の 1 つは，時間間隔を考慮して特徴ベクトルを補正することである．つまり，時間間隔の大きさに伴い，同義性を持つデータ対であっても，その特徴ベクトルのある次元の要素の値が小さくなる(あるいは大きくなる)場合に，その値を補正して同義性を持つデータ対の典型的な特徴ベクトルに近づけることである．これは，Li らの時間減衰に近い方式である[Li et al. 2011]．しかし，その補正のモデルは人手，あるいは機械学習で与える必要があり，準備のコストが大きく実用性に欠ける問題がある．

(2) 時間間隔を尺度の一部として採用する方法

もう 1 つの手法は，既存の能動学習のクエリ選択基準に時間間隔の尺度を組み合わせる方式である．両尺度共に事前の定義が容易であり，実用性が高い．

ラベル無しデータ集合$\mathcal{U}$があり，あるモデルθに対するあるデータ対$x \in \mathcal{U}$の確信度を$\mathrm{c}(x,\theta)$，データ対xの時間距離を$\mathrm{t}(x)$とする．選択すべきデータ対をx^*と表記すると，この方式によるクエリ選択は以下のように定式化できる．

$$x^* = \underset{x \in \mathcal{U}}{\operatorname{argmin}} (1-\alpha) * \mathrm{c}(x,\theta) + \alpha * \mathrm{t}(x) \qquad (3.1)$$

ここで，α $(0 \le \alpha \le 1)$は選択尺度における時間距離の比重である．$\mathrm{c}(x,\theta)$は既存研究の様々な尺度を採用することができ，例えば SVM 超平面に対するユークリッド距離によって算出できる．$\mathrm{t}(x)$はデータ対xの時間間隔に対して単調増加である，任意の関数である．

4. 評価

4.1 実験設定

(1) 利用データ

DBLP データにおける Hiroyuki Sato の名前を持つ著者の変遷を伴う同義性を判定する実験を行う．Hiroyuki Sato，Sato Hiroyuki の名前を持つ 57 のレコードのうち，55 のレコードについて実在の人物に同定する(上記レコードは，11 の実在の人物に人手で分類した．2 つのレコードは人物の特定が不可能であったため，除外している)．所属組織情報は，元の Hiroyuki Sato に関する論文から人手で抽出し，論文に関する技術用語情報は，著者キーワード，及び抄録から特徴語を抽出して利用する．

(2) 実装

機械学習方式は SVM を採用する．3.1 式の確信度$\mathrm{c}(x,\theta)$はデータ対xの特徴ベクトルの，超平面θまでのユークリッド距離として実装する．また，時間距離$\mathrm{t}(x)$は以下のように定義する．

$$\mathrm{t}(x) = \beta * \Delta\mathrm{t}(x)^{\gamma} \qquad (4.1)$$

関数$\Delta\mathrm{t}(x)$は，データ対xの時間間隔を求める関数，βとγは正の実数である．本実験では$\beta = 0.1$, $\gamma = 1$と設定する．

本実験では，3 つのクエリ選択基準に対して比較評価を行う．1 つは提案方式である時間距離を考慮した基準(以下，TIME)であり，他は先行技術である時間距離を考慮しない基準(以下，CERT．3.1 式において$\alpha = 0$の場合)，及びランダムに選択する基準(以下，RAND)である．

本実験では，データセットの実世界での実体のうち，半分を同義性判定モデルの学習に用い，もう半分のデータ群をテストデータとしてそのモデルを適用する，交差検定の方式で上記 3 つの方式の評価を行う．本実験データの場合，実在の人物の 11 の実体のうち，5 つの実体をモデル学習用として用い，残りをテストデータとして用いる．学習用として選ばれる実体のパターンは${}_{11}\mathrm{C}_5 = 462$通りあるが，ランダムに 50 通り選択する．

実験のフローは以下の通りである．最初の「ラベル付き教師データ集合」としてランダムに 4 つのデータ対を選択し，最初の機械学習モデルを生成する．また，モデルに対する評価値を算出する．そのモデルに対して各基準に基づきクエリとして与えるデータ対を 2 つ選択し，モデルを再学習させ，評価値を算出する．本実験では，この手順を 40 回繰り返す．

(3) 評価基準

変遷を伴う同義性判定問題には，クラスタリング技術の評価基準が適切である．本研究では正規化相互情報量(normalized mutual information, 以下 NMI) [Zhong and Ghosh 2005]を評価に用いる．正規化相互情報量では 0 から 1 の間の実数の値が得られ，クラスタリング精度が良いほど高い値が算出される．

実験を行う 50 通りの教師データのパターンはそれぞれ，そのパターンから得られる最適な機械学習モデルθ^*を用いてテストデータを判定する際の最良の精度(精度の限界値)が大きく異なる．本実験では，算出した NMI の値を各教師データのパターンで得られる精度限界の NMI 値で正規化した値(nNMI と表記する)を，各手法を比較する評価値として用いる．精度限界 NMI の推定値は，教師データ用の 5 つの実体から，2.3 節(3)の戦略に従って得られる判定モデルθ^*を採用した場合の NMI 値とする．即ち，あるモデルθをテストデータ集合Tに適用した時の NMI 値を$\mathrm{NMI}(\theta, \mathrm{T})$とすると，nNMI 値は以下の式で算出される．

$$\mathrm{nNMI}(\theta, \mathrm{T}) = \frac{\mathrm{NMI}(\theta, \mathrm{T})}{\mathrm{NMI}(\theta^*, \mathrm{T})} \qquad (4.2)$$

但し，あくまで推定であるため，nNMI は 1 を超える場合がある．

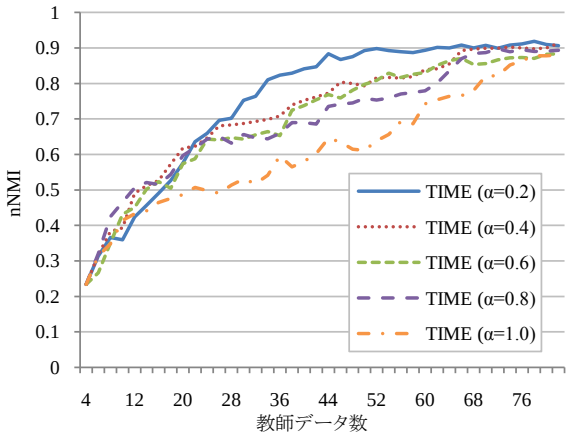


図 4:時間距離を考慮した能動学習(TIME)の結果

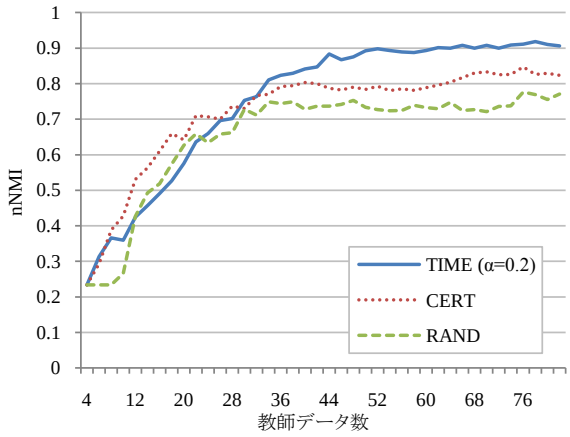


図 5:3 つの方式による能動学習結果の比較

4.2 実験結果

図 4 は TIME の方式において，時間距離の比重αの違いによる実験結果を示す．各実験結果は，50 通りの教師データのパターンで得られた nNMI 値の平均値を算出している．α = 0.2～0.8の場合は，教師データ数が 26 個程度までは精度に大きな違いは見られないが，それ以降はα = 0.2以外の場合の改善の傾向は穏やかになり，違いが明確に表れている．なお，α = 1の場合，即ち時間距離のみを考慮する場合は教師データ数が 16 辺りから，精度改善の度合いが弱まっている．但し，いずれのパラメータの場合も，早さに違いはあるが，nNMI 値は 0.9 程度まで改善する点が特徴的である．

図 5 は 3 つのクエリ選択基準で追加した教師データの数に対する平均の nNMI 値の変化を示す．教師データ数が 32 付近までは，3 つの方式共にほぼ同程度の精度であるが，それ以降は CERT, RAND の方式の改善はほぼ停止し，TIME の方式のみが継続して改善している．TIME の方式が nNMI 値 0.9 まで改善するのに対し，CERT は 0.8, RAND は 0.75 程度の精度で改善がほぼ停止している．

5. 考察

実験結果より，変遷を伴う同義性判定問題に能動学習の方式を適用する場合，時間距離の小ささを考慮したクエリ選択基準により，従来の不確実さに基づく基準に対して最終的に nNMI 値で 0.1 ポイント程度高い精度の能動学習が実現できることが分かった．しかし一方で，nNMI 値は 0.9 程度で精度向上は停止した．本アプローチの一定の有効性は示せたが，3.2 節で示した「ノイズとなるデータ対の選択を避けること」という課題に対してはまだ取り組むべき点があると考えられる．

また，図 4 より，時間距離を考慮したクエリ選択尺度は最終的にはパラメータαに依存することなく同程度の精度になったが，その早さはαの値に大きく依存することが分かった．改善のピークに早く到達するという観点では不確定度に基づく尺度が優れており，精度を高めるという観点では時間距離の尺度が優れていると本実験結果から推察される．

6. おわりに

本稿では変遷を伴う同義性判定問題に対して能動学習を効率的に実施する方式として，従来の不確実性に基づく尺度に時間距離の小ささという尺度を加えることを提案した．実験により，本提案が精度改善のピーク値を高め，また時間距離の小ささの尺度を考慮する比重のパラメータを適切に設定することで，少ない教師データ数で精度を大きく向上させることができることを示し，本アプローチの一定の有効性を明らかにした．しかし，今回は研究の開始段階として，本問題に対して時間距離の考慮の有効性を示すことに重点を置いているため，能動学習の最適性までは考慮していない．今後，先行研究の様々な知見を活かし，より高い精度により少ない教師データで到達できるような，最適な能動学習方式を追求する予定である．

参考文献

[Bizer et al. 2009] Bizer, C., Heath, T., and Berners-Lee, T.: “Linked Data - The Story So Far,” International Journal on Semantic Web and Information Systems, Special Issue on Linked Data, 2009.

[Elmagarmid et al. 2007] Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S.: “Duplicate Record Detection: A Survey,” IEEE Trans. on Knowl. and Data Eng., vol.1, no.19, pp.1-16, 2007.

[Li et al. 2011] Li, P., Dong, X. L., Maurino, A., and Srivastava, D.: “Linking Temporal Records,” In Proceedings of the VLDB Endowment, vol.4, no.11, pp. 956-967, Aug. 2011.

[Mirroshandel et al. 2011] Mirroshandel, S. A., Ghassem-Sani, G. R., and Nasr, A.: “Active Learning Strategies for Support Vector Machines, Application to Temporal Relation Classification,” In Proceedings of International Joint Conference on Natural Language Processing (IJCNLP 2011), pp.56-64, 2011.

[Schohn and Cohn 2000] Schohn, G. and Cohn, D.: “Less is More: Active Learning with Support Vector Machines,” In Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00), pp.839-846, 2000.

[Settles 2010] Settles, B.: “Active Learning Literature Survey,” Technical Report 1648, Univ. of Wisconsin-Madison, 2010.

[Weikum et al. 2011] Weikum, G., Ntarmos, N., Spaniol, M., Triantafillou, P., Benczur, A., Kirkpatrick, S., Rigaux, P., and Williamson, M.: “Longitudinal Analytics on Web Archive Data: It’s About Time!” In Conference on Innovative Data Systems Research, pp.199-202, 2011.

[Zhong and Ghosh 2005] Zhong, S., and Ghosh, J.: “Generative model-based document clustering: a comparative study,” Knowl. Inf. Syst., vol.3, no.8, pp.374-384, 2005.

[森田他 2011] 森田大翼，飯塚京士，村山隆彦，赤埴淳一: “ナレッジ活用のための機械学習による変遷情報の意味論的統合，” 信学技報，vol.111, no.310, AI2011-20, pp.19-24, 2011.