DEIM Forum 2017 F6-5

文脈誘導型ランキング学習

加藤　誠† 内田　臣了† Wiradee　Imrattanatrai† 山本　岳洋† 大島　裕明†
田中　克己†

† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †{kato,uchida,wiradee,tyamamot,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　本論文では，順序付けられた一部のエンティティを訓練データ，エンティティの数値属性を特徴として線形関数を学習することで，エンティティのランキングを行う問題に取り組む．例えば，$f(\text{Entity}) = +0.5(\text{GDP}) - 0.8($ 自殺者数 $)$ という関数を「幸福度」によって順序づけられたエンティティから学習する．この問題では訓練データの数が非常に少ないため，Web 上に存在する大量の「文脈」による補助を受けて学習を行う機械学習手法，文脈誘導型学習を本論文では提案する．実験では 3 クラスのエンティティについて 158 個の様々な順序によるランキングを行った．実験結果から，既存のランキング学習手法よりも高い精度が文脈誘導型学習によって得られることが明らかとなった．
キーワード　ランキング学習，エンティティランキング，Web マイニング

1. はじめに

エンティティの順序は人気の高いコンテンツであり，多くの Web サイトで紹介されている．例えば，国，都市，企業，有名人，および，製品などが，居住性，革新性，美しさ，性能などのさまざまな基準によって順序づけられている．順序は比較文（例えば，“レオナルド ディカプリオはブラッド ピットよりも背が高い”）や，最上級（例えば，“東京は最も物価の高い都市である”），ランキング（例えば，“1 位：デンマーク，2 位：スイス，3 位：アイスランド”）などの形式で表現されている．人々はそのようなエンティティの順序を意思決定，比較，計画などに利用する．

順序づけの基準の詳細は，場合によっては正確かつ客観的に記述されているが，省略されたり基準が主観的である場合もある．例えば，デンマークはいくつかのウェブサイトで最も幸せな国として紹介されているが詳細な証拠が提示されていない場合も多い．その根拠は，国民の自己申告，他の国の評判，あるいは，GDP や自殺者数などの客観的尺度かもしれない．不確かな証拠だけでそのような知識を獲得することは，偏見やステレオタイプにつながる可能性がある．

本論文では，部分的に観測された順序を訓練データとして，エンティティの数値属性を特徴として使用することで，エンティティの順序を学習する方法を提案する．学習されたモデルによって，Web 上の順序を説明することが可能であり，また，学習に使用されていない他のエンティティをランキングすることも可能である．順序付けは数値属性の線形関数 f によって表すことができると仮定する．これは，主にユーザへの説明能力が高く，現実世界における順序の大部分は，いくつかの尺度を線形結合することによって求められるという事実があるためである．例えば,「幸福度」という「順序基準」によって順序づけられた（デンマーク，スイス，アイスランド）というエンティティ列と「GDP」,「国土面積」,「自殺者数」という数値属性が与えられた時，$f(\text{Entity}) = +0.5(\text{GDP}) - 0.8($ 自殺者数 $)$ という線形関数を学習することができる．

この問題の主な課題の 1 つは，訓練データが不足していることである．多くの Web サイトでは，すべてのエンティティの順序は記述されてはいない．さらに，一部のエンティティクラスでは，すべてのエンティティの順序が記述されていたとしても訓練データのサイズは十分ではない（たとえば，日本の都道府県数は 47 であり，アメリカの州の数は 50 である）．数値属性の数は，通常，Web 上で順序づけられたエンティティの数よりもはるかに多いので，エンティティの順序を学習する際に，深刻な過学習の問題が起こりうる．

この本質的な問題に対処するために，我々は**文脈誘導型学習**と呼ばれる学習方法を提案する．文脈誘導型学習は，順序づけられたエンティティだけでなく，各特徴の重みを学習するために順序基準と属性に関する文脈を利用する．文脈は追加の情報をモデルに与え，過学習を防止して学習を正しい方向に誘導する．例えば，“幸福度” という基準に関する線形関数の，“GDP” という特徴に対する重みは,（デンマーク，スイス，アイスランド）というエンティティ列だけでなく，“**GDP** は**幸福**のための重要な要素である” や “**GDP** が向上すれば**幸福度**も改善する”，といった「幸福」と「GDP」の文脈（「幸福」と「GDP」の関係について記述された文や文書）によっても推測されることになる．文脈誘導型学習は，文脈に対するアノテーションを必要とせず，代わりに，文脈と関数 f 中の重みの関係を学習するために，複数の順序を同時に学習する．

我々の知る限りにおいて，文脈誘導型学習は，機械学習において「ラベル基準」と特徴の文脈を直接活用する最初の試みである（ラベル基準とはラベル付けの基準を表すテキストのことであり，エンティティランキングの問題における順序基準に対応する）．文脈誘導型学習は，ラベル基準と特徴の関係が特定のコーパスに記述されてれば，順位付けだけでなく分類や回帰の問題にも適用することが可能である．

実験では，国，都道府県，カメラの 3 つのエンティティクラスを対象とし 158 種類の順序を学習した．訓練された順序の精度を測定し，与えられた順序が学習されたモデルによってどれだけ説明可能であるかを評価した．実験の結果，文脈誘導型学習は既存のランキング学習手法よりも統計的有意に高い精度を示した．

この論文における我々の貢献を以下に示す：(1) 数値属性を特徴とした，エンティティの順位付けを学習する問題を提案した．これによって，Web 上の順序を正確に理解し，さまざまな基準でエンティティを順序付けすることが可能である．(2) 文脈誘導型学習を提案した．これは，過学習を防止するためにラベル基準および特徴の文脈を利用する一般的な機械学習モデルである．(3) 幅広い順序に対する実験を行い，エンティティの順序づけタスクにおいて文脈誘導型学習の有効性を示した．

本論文の構成は以下の通りである．2 節ではエンティティのランキングに関する関連研究，および，文脈誘導型学習と他の機械学習手法との関係について述べる．3 節では問題設定を説明し，文脈誘導型学習，および．その主問題への適用方法について述べる．4 節では実験結果を示す．最後に，5 節では今後の課題と共に本論文の結論を述べる．

2. 関 連 研 究

本節では，エンティティのランキングに関する関連研究，および，文脈誘導型学習と他の機械学習手法，特にマルチタスク学習との関係について述べる．

2.1 エンティティランキング

エンティティランキングは INEX や TREC のいくつかのトラックで取り組まれてきている．INEX の Entity Ranking トラックでは，エンティティランキングとエンティティリスト補完の 2 つのタスクが提案された [14–16]．エンティティランキングタスクは，与えられたクエリに対して適合するエンティティを発見するタスクであり，他方，エンティティリスト補完タスクでは与えられたエンティティ例に関連するエンティティを発見するタスクであった．TREC のエンティティトラックでは，あるエンティティ例と発見対象のエンティティタイプ，それらの関係性が与えられたときに関連するエンティティを取得する，関連エンティティ発見タスクが提案された [2–4]．これらのタスクでは，発見されたエンティティが与えられたエンティティ例との関連性に基づいて順序づけされることだけが求められ，関連エンティティに対して様々な順序付けを行うことまでは求められていない．

エンティティリスト補完タスクと関連エンティティ発見タスクとは別に，エンティティのランキング学習に関する関連研究も存在する．Kang らは与えられたクエリに関連するエンティティを発見するために，決定木ブースティングに基づくランキングアルゴリズムを用いている [23]．Tran らはイベントのタイムライン要約のために主要度と情報量に基づいてエンティティをランキングする手法を提案している [32]．Zhou らは与えられたエンティティと特定の関係にあるようなエンティティを発見する問題に取り組んでいる [35]．例えば，ユーザは “FounderOf” という関係と “Microsoft” というエンティティを入力し “Bill Gates” を得ることができる．彼らはあるエンティティ対の関係に関する検索スニペットから得られた特徴を用いて，訓練クエリとラベル付けされたエンティティを元に各関係についてランキングを学習した．この研究と我々の研究は共に文脈（または検索スニペット）を用いてランキング学習を行うという点では類似するが，我々のモデルは主にエンティティの属性に基づくものであり，エンティティ対の文脈は用いていない．

我々のタスクに類似した自然言語処理タスクもいくつか提案されている．Iwanari らは，与えられた形容詞に基づいてエンティティを順序づける問題に取り組んでいる [21]．手法はテキストから得られたいくつかの根拠に基づいており，エンティティと形容詞の共起やエンティティと形容詞の依存関係，直喩，比較文などが用いられている．このタスクには既存のランキング学習および回帰モデルが用いられている．ある順序基準に基づいてエンティティをランキングを行うという点でこのタスクは我々のタスクと類似する．彼らの手法は順序基準とエンティティの文脈を利用するが，我々の手法では順序基準とエンティティの「属性」の文脈を利用している．また，本論文ではこのタスクについて効果的な学習手法についても提案を行っている．エンティティの数値的な属性を抽出する研究も存在するが，我々のタスクでの順序基準は多くの場合に定量化することが困難であり，これらの研究で提案される手法では推定できないと思われる [13, 27, 31]．

2.2 マルチタスク学習

文脈誘導型学習の重要な特徴を以下に要約する：(1) 関数 f の重みはラベルだけでなく，ラベル基準と特徴の文脈に基づいて学習される，また，(2) 文脈と関数 f 中の重みの関係を学習するために，複数の関数が同時に学習される．以下では，いくつかの機械学習手法について述べ文脈誘導型学習との関連について議論する．

マルチタスク学習とは複数のタスクを同時に学習することによって個々のタスクの学習を改善する方法である [10]．文脈誘導型学習はマルチタスク学習の一種であると考えられる．Evgeniou と Pontil によって提案された正則化マルチタスク学習は複数のタスクにおける関数の重みが似ていることを仮定している [17]．後に述べるように，彼らのモデルは我々のモデルの特殊な場合であり，すべての文脈が同じである場合に等価なモデルとなる．また，共通の事前分布から重みが生成されることを仮定するモデルもある [12, 26, 34]．Argyriou らは重みが複数のタスクに共通する低次の部分空間において表現できることを仮定している [1]．これらのように全てのタスクが関連すると仮定する研究とは対照的に，いくつかの研究ではどのタスクが関連し，似た重みを共有すべきか選択するようなモデルを提案している [22, 24]．同様に，文脈誘導型学習はタスク間の類似性を文脈を用いることで暗に推定し，似たようなタスクに対して似たような重みを推定する傾向にある．文脈誘導型学習と他のマルチタスク学習手法の興味深い違いは，いずれのタスクも類似しない場合においても文脈誘導型学習は適用可能であるという点である．文脈誘導型学習は複数のタスク間でいくつかの文

脈が類似することだけを仮定する．したがって，文脈誘導型学習の適用範囲は既存のマルチタスク学習が対象とする問題に限らない．

最後に，文脈誘導型学習（Context-guided Learning，CGL）と似た名前を持つ機械学習手法との差違を明確にする．文脈依存型学習（Context-*sensitive* Learning）はしばしば特徴の文脈（例えば，ピクセルを特徴とする場合の周辺ピクセル [5] や単語を特徴とする場合の共起単語 [11] など）を用いる手法を指す．他の場合，与えられたデータに対して選択的にモデルを用いる機械学習手法 [30] や直前・直後のデータを予測に用いる手法 [28] を文脈依存型学習と呼ぶ．これらの手法は主に文脈を追加の特徴として用いることに焦点を当てており，文脈誘導型学習はモデルを変えることなく，文脈を用いてモデルの学習を改善する手法である．さらに，文脈誘導型学習で用いられる文脈は特徴に関するものではなく，ラベル基準と特徴に関するものである．

3. 手　　法

本節では，まず部分的に観測されたエンティティの順序から，エンティティの順序を数値属性に基づいて学習する問題について説明を行う．それから，文脈誘導型学習を導入し，その主問題への適用について述べる．

3.1 問 題 設 定

E をあるクラスのエンティティ集合とし，エンティティ順序 $\preceq_k$ を E 上の全順序として定義する．各順序は順序の基準を示す順序基準 l_k を持つ．例えば，順序基準として「住みやすさ」，「革新性」,「美しさ」,「性能」などが挙げられる．順序付き集合 $O_{\preceq_k}$ $(\subset E \times E)$ は $e_i \preceq_k e_j$ を満たす全てのエンティティ対の集合 $(e_i, e_j) \in E \times E$ である．Web 上においてエンティティ順序はこれら順序付き集合の部分集合によって表現されることが多い．従って，エンティティ順序の学習のために観測・利用可能なのは $O'_{\preceq_k} \subset O_{\preceq_k}$ である．例えば，“レオナルド ディカプリオはブラッド ピットよりも背が高い” という文は $O'_{\preceq_k} = \{$(“ブラッド ピット”, “レオナルド ディカプリオ”)$\}$ を示し，“1 位：デンマーク，2 位：スイス，3 位：アイスランド” というランキングは $O'_{\preceq_k} = \{$(“スイス”, “デンマーク”), (“アイスランド”, “デンマーク”), (“アイスランド”, “スイス”)$\}$ を示す．

我々の主目的は，各エンティティ $\preceq_k$ に対して順序部分的に観測された順序付き集合 $O'_{\preceq_k}$ から線形関数 $f_k(e_i) = \mathbf{w}_k^T \mathbf{e}_i$ を学習することである．ただし，$\mathbf{e}_i$ はエンティティ $e_i \in E$ の数値属性を表す M 次元ベクトルであり，このベクトルの d 次元目はエンティティの a_d という属性を表す．我々はこの関数 f_k がエンティティ順序 $\preceq_k$ を**保つ**ことを期待する：つまり，任意の $e_i, e_j \in E$ に対して $e_i \preceq_k e_j \Rightarrow f_k(e_i) \leqq f_k(e_j)$ を満たすことを期待する．学習された関数 f_k を用いることで順序基準 l_k によってエンティティを順位付けすることが可能であり，学習された関数中において 0 でない重みを持つ数値属性によって順序基準を説明することができる．

先に述べたように，この問題の主となる課題は訓練データの不足である．より正確には，数値属性の数 M と比較して $|O'_{\preceq_k}|$ がしばしば小さいことである．例えば，我々の実験において「国」では $M = 83$,「都道府県」では $M = 137$ である．10 件以下のエンティティしかランキングされていない場合，訓練データとして我々が得られるのは高々45 個のエンティティペアであり，これらのクラスにおいては学習に十分な量であるとは考えにくい．さらに，幅広い順序を扱うためには M はできるだけ大きい方が良い．したがって，訓練データの不足から起こりうる過学習を防ぐためになんらかの手段が必要となる．

本研究における主なアイデアは $\mathbf{w}_k$ の学習のために，$O'_{\preceq_k}$ 以外のデータを活用することである．我々の問題の特徴，または，仮定として，順序基準と属性に対するテキスト表現が利用可能であることが挙げられる．そのため，順序基準と属性に関する文脈を活用することによって，各属性に対する重みをおおよそ推定することができるのではないかと我々は考えた．例えば，“**GDP** が向上すれば**幸福度**も改善する” という文脈からは「GDP」に対する正の重みを，“**自殺者数**が増えれば**幸福度**が低下する” という文脈からは「自殺者数」に対する負の重みを推定する．一方で，ある順序基準と属性の対に対して，もし文脈が得られなかった場合には，その属性に対する重みは 0 に近いという推測をする．これらのアイデアは次節で説明する文脈誘導型学習によって具体化される．

3.2 文脈誘導型学習

本節ではラベル基準と特徴の文脈を活用する文脈誘導型学習という学習手法を提案する．まず，分類問題に対する文脈誘導型学習を導入し，次にランキング問題に対する拡張を行う．

入力はベクトルとラベルのペアの集合の集合である：$\mathcal{D} = \{D_k\}_{k=1}^K$．ただし，$D_k = \{(\mathbf{x}_{k,i}, y_{k,i})\}_{i=1}^{N_k}$，$\mathbf{x}_{k,i} \in \mathbb{R}^M$，$y_{k,i} \in \{-1, +1\}$，そして，$K$ はラベル基準の総数である．ラベル基準 l_k はラベル $y_{k,i}$ を決定するためのテキスト表現である．例えば，$\mathbf{x}_{k,i}$ が都市の特徴を表し，都市が大都市であれば $y_{k,i} = +1$，そうでなければ $y_{k,i} = -1$ とする．この場合，ラベル基準 l_k は「大都市」となる．例えば，$\mathbf{x}_{k,i}$ が文書の特徴を表し，文書がスパムであれば $y_{k,i} = +1$，そうでなければ $y_{k,i} = -1$ とする．この場合，ラベル基準 l_k は「スパム」となる．ベクトルの d 次元目の値はある特徴に対応し，その特徴の名前を a_d とする．例えば，特徴名として「人口」や「面積」，「リンク数」などが考えられる．

文脈誘導型学習を適用するための要件は以下の 3 つである：

（1） ラベル基準 l_k が言語で表現されている．
（2） 特徴 $A = \{a_d\}_{d=1}^M$ が言語で表現されている．
（3） ラベル基準と特徴の関係に関する文脈を含むコーパスが存在する．

マルチタスク学習問題とは異なり，文脈誘導型学習ではタスクの類似性を仮定しない（または，文脈誘導型学習の言葉で言えば，ラベル基準が類似する必要がない）．また，全てのラベル基準および特徴が言語で表現されている必要もない．

分類問題は，各ラベル基準 $k = 1, \ldots, K$ に対して $f_k(\mathbf{x}_{k,i}) \simeq y_{k,i}$ であるような関数 f_k を学習する問題であると定式化できる．文脈誘導型学習によってこの問題を解くために，我々は

線形関数 $f_k(\mathbf{x}_{k,i}) = \mathbf{w}_k^T \mathbf{x}_{k,i}$ を用いる．ラベル基準 l_k と特徴 $a_d \in A$ の文脈を $c_{k,d}$ とし，以下のようにして $\mathbf{w}_k$ を推定するために文脈を用いることができる：

$$w_{k,d} = \mathbf{u}^T \phi(c_{k,d}) + v_{k,d}, \tag{1}$$

ただし，$w_{k,d}$ は $\mathbf{w}_k$ の d 次元目の値，ϕ は文脈からベクトルへの特徴写像関数，$\mathbf{u}$ はラベル基準に依存しない重みベクトルである．上記の式は，あるラベル基準 l_k と特徴 a_d の重みがそれらの文脈である $c_{k,d}$ と切片 $v_{k,d}$ から推定されることを意味する．また，我々は $v_{k,d}$ が「大きくない」ことを，つまり，$w_{k,d}$ が $v_{k,d}$ のみによって決定されるのでなく，$\mathbf{u}^T\phi(c_{k,d})$ によっても決定されることを期待する．式 1 は正則化マルチタスク学習 [17] における $w_{k,d} = z_d + v_{k,d}$（z_d は複数タスクに共通の重み）を一般形である．もし全ての文脈が同じであれば式 1 は正則化マルチタスク学習に還元される．もし 2 つのラベル基準の文脈が似る，つまり，ラベル基準が類似するのであれば，$w_{k,d}$ はそれらのラベル基準に対して似た値となる．この性質はいくつかのマルチタスク学習手法に似ている [22, 24].

本論文ではサポートベクターマシン（SVM）と正則化マルチタスク学習 [17] に似た，正則化によるアプローチによって，文脈を含む線形関数を学習する．重みを学習するための最適化問題を以下に示す：

Problem 1.

$$\min_{\mathbf{u},\mathbf{v}_k,\xi_{k,i}} \|\mathbf{u}\|^2 + \frac{c}{K}\sum_{k=1}^{K}\|\mathbf{v}_k\|^2 + C\sum_{k=1}^{K}\sum_{i=1}^{N_k}\xi_{k,i}, \tag{2}$$

subject, for $k = 1,\ldots,K$ *and* $i = 1,\ldots,N_k$*, to the constraints that*

$$y_{k,i}\mathbf{w}_k^T\mathbf{x}_{k,i} \geqq 1 - \xi_{k,i}, \tag{3}$$

$$\xi_{k,i} \geqq 0, \tag{4}$$

where c *and* C *are parameters,* $\mathbf{v}_k = (v_{k,1},\ldots,v_{k,M})$*,* $\mathbf{w}_k = \Phi_k^T\mathbf{u} + \mathbf{v}_k$*, and* $\Phi_k = (\phi(c_{k,1}),\ldots,\phi(c_{k,M}))$.

スラック変数 $\xi_{k,i}$ は訓練データにおける線形関数の誤差を表し，他の 2 つの項は $\mathbf{u}$ と $\mathbf{v}_k$ に対する正則化項である．パラメータ c と C はそれぞれ文脈のモデルに対する影響と訓練データにおける誤差への感度を調節する役割を持つ．パラメータ c に対する高い値は文脈の影響を高め，一方でパラメータ C に対する高い値は訓練データにおける誤分類を許容しにくくさせる．関数 f_k は同時に学習するのではなく，個別に学習することも可能ではある．しかし，その場合，利用可能なデータ量を増やすことなく，単に推定すべき重みの数を増加させてしまう．

次に，Problem 1 が一般的な SVM と同じように解けることを示す．このために，まず関数 f_k $(k = 1,\ldots,K)$ を要約するような，学習されるべき 1 つの関数を以下のように定義する：

$$F(\mathbf{x}, k) = f_k(\mathbf{x}) \tag{5}$$

この関数 $F : \mathbb{R}^M \times \{1,\ldots,K\} \to \mathbb{R}$ は以下の線形関数として表現することができる：

$$F(\mathbf{x}, k) = \mathbf{w}^T\psi(\mathbf{x}, k) \tag{6}$$

ただし，

$$\mathbf{w} = (\sqrt{\frac{K}{c}}\mathbf{u}^T, \mathbf{v}_1^T, \ldots, \mathbf{v}_K^T)^T, \tag{7}$$

かつ，

$$\psi(\mathbf{x}, k) = (\sqrt{\frac{c}{K}}(\Phi_k\mathbf{x})^T, \underbrace{\mathbf{0}^T,\ldots,\mathbf{0}^T}_{k-1}, \mathbf{x}^T, \underbrace{\mathbf{0}^T,\ldots,\mathbf{0}^T}_{K-k})^T \tag{8}$$

である．ここで，ψ は特徴写像関数，$\mathbf{0}$ は要素がすべて 0 であるような M 次元ベクトルである．

このとき，Problem 1 は以下に示すように一般的な SVM の問題へと帰着できる．

Theorem 1. *The optimization of Problem 1 is equivalent to solving the following problem:*

Problem 2. *Given* $D = \{((\mathbf{x}_i, k_i), y_i)\}_{i=1}^N$ *where* $N = \sum_{k=1}^K N_k$ *such that* $D = \bigcup_{k=1}^K \{((x_{k,i}, k), y_{k,i}) | (x_{k,i}, y_{k,i}) \in D_k\}$*,*

$$\min_{\mathbf{w},\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C'\sum_{i=1}^{N}\xi_i, \tag{9}$$

subject, for $i = 1,\ldots,N$*, to the constraints that*

$$y_i\mathbf{w}^T\psi(\mathbf{x}_i, k_i) \geqq 1 - \xi_i, \tag{10}$$

$$\xi_i \geqq 0, \tag{11}$$

where $C' = \frac{CK}{2c}$ *and* ξ_i *is a slack variable for* $((x_i, k_i), y_i) \in D$*.*

Proof. $i = \sum_{k'=1}^{k-1} N_{k'} + i'$ として，Problem 2 の ξ_i に $\xi_{k,i'}$ を，$\mathbf{x}_i$ に $\mathbf{x}_{k,i'}$ を，y_i に $y_{k,i'}$ を代入すれば，Problem 1 を得られる． □

Problem 2 は一般的な SVM の問題であるため，以下に示すように一般的な SVM の双対問題を利用することで Problem 1 を解くことができる．

Theorem 2. *Define a kernel function* $K_{k,k'}$ *as:*

$$\begin{aligned} K_{k,k'}(\mathbf{x}_i, \mathbf{x}_{i'}) &= \psi(\mathbf{x}_i, k)^T\psi(\mathbf{x}_{i'}, k'), \\ &= \frac{c}{K}\mathbf{x}_i^T(\Phi_k^T\Phi_{k'} + \delta_{k,k'}\mathbf{I})\mathbf{x}_{i'}, \end{aligned} \tag{12}$$

where $\delta_{k,k'} = 1$ *if* $k = k'$*, and* 0*, otherwise.*

The dual problem of Problem 2 is given by:

Problem 3.

$$\max_{\alpha_i} \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i y_i \alpha_i y_{i'} K_{k_i,k_{i'}}(\mathbf{x}_i, \mathbf{x}_{i'}), \tag{13}$$

subject, for $i = 1,\ldots,N$*, to the constraint that*

$$0 \leqq \alpha_i \leqq C'. \tag{14}$$

If α_i^* *is a solution for the problem above, then the solution to Problem 1 is given by*

$$f_k^*(\mathbf{x}_{k,i}) = \sum_{i'=1}^{N}\alpha_{i'}^* y_{i'} K_{k_{i'},k}(\mathbf{x}_{i'}, \mathbf{x}_{k,i}). \tag{15}$$

上記の双対問題を解くためには，N 個の訓練データを用いて通常の SVM を解くのと同じ計算量が必要となる（N は全てのラベル基準に対するベクトルとラベルのペアの総数であり，$N = \sum_{k=1}^{K} N_k$ である）．例えば，n 個の訓練データに対して通常の SVM を解くのに $O(n^3)$ の計算量が必要であれば，上記の双対問題を解くのにかかる計算量は $O(N^3)$ となる．一方，通常の分類問題のように異なるラベル基準に対して別々の関数を学習するような場合には，$O(\sum_{k=1}^{K} N_k^3)$ の計算量のみが必要となる．本論文中で行う実験ではラベル基準の数が 40 から 64 程度であるため計算量が大きな問題とはならないが，特にラベル基準が多いとき，計算量は文脈誘導型学習の欠点となり得る．

我々の問題は一般的な SVM の問題に還元できるため，SVM と同様にカーネルによって非線形関数を扱うことができる．モデルの性質上，分類のための線形関数 f_k は非線形関数にすることはできないが，文脈に基づく重みの推定（式 1 参照）のために非線形関数を用いることが可能である．Problem 3 にて見られるように，最適な関数を求めるためには文脈から得られるベクトル $\phi(c_{k,d})$ のカーネル関数のみが必要であることがわかる．したがって，特徴写像関数 ϕ をここで $\phi : \mathcal{C} \to \mathcal{H}$ と定義できる．ただし，$\mathcal{C}$ は文脈集合，$\mathcal{H}$ は可分な一般抽象ヒルベルト空間である．この特徴写像関数 ϕ に対応するカーネルは以下のように定義される：

$$\kappa(c_{k,d}, c_{k',d'}) = \langle \phi(c_{k,d}), \phi(c_{k',d'}) \rangle_{\mathcal{H}}, \tag{16}$$

ただし，$\langle \cdot, \cdot \rangle_{\mathcal{H}}$ は $\mathcal{H}$ 中での内積を表す．

Problem 3 中のカーネル関数 $K_{k,k'}$ を以下のように置き換える：

$$K_{k,k'}(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{c}{K} \mathbf{x}_i^T (\Phi_{k,k'} + \delta_{k,k'} \mathbf{I}) \mathbf{x}_{i'}, \tag{17}$$

ここで，$\Phi_{k,k'}$ の (d, d') 要素は $\kappa(c_{k,d}, c_{k',d'})$ である．上記の置き換えによって，カーネル κ の計算だけが必要となり，ϕ による特徴写像の計算が不要となる．

文脈誘導型学習におけるカーネルの利用によって，幅広い文脈の表現を行うことが可能となる．文脈のカーネル関数 κ が適切に設計されている限り，ベクトル，ベクトル集合，グラフなど様々な文脈表現が可能である．

3.3 文脈誘導型ランキング学習

ここでは，文脈誘導型学習をランキング問題へと拡張し，エンティティランキングの問題に適用する方法について説明する．

ランキング問題の入力は $\mathcal{D} = \{D_k\}_{k=1}^{K}$ であり，$D_k \subseteqq \mathbb{R}^M \times \mathbb{R}^M$，$K$ は順序基準の数である．順序基準 l_k は D_k の順序を決定するためのテキスト表現であり，D_k において $(\mathbf{x}_{k,i}, \mathbf{x}_{k,j})$ であることは，$\mathbf{x}_{k,j}$ が順序基準 l_k の観点から $\mathbf{x}_{k,i}$ に勝ることを示す．D_k のベクトルにおける d 次元目の値はある特徴と対応しており，その特徴の名前を a_d とする．文脈誘導型ランキング学習の要件は，分類問題のための文脈誘導型学習と同じである．ランキング問題は各順序基準 $l_k (k = 1, \ldots, K)$ において，D_k 中の $(\mathbf{x}_{k,i}, \mathbf{x}_{k,j})$ に対して $f_k(\mathbf{x}_{k,j}) - f_k(\mathbf{x}_{k,i}) \simeq 1$ となるような関数 f_k を学習することであると定式化できる．分類問題においても仮定されていたように，線形関数 $f_k(\mathbf{x}_{k,i}) = \mathbf{w}_k^T \mathbf{x}_{k,i}$ を学習するものとする．

以下のように D_k を再定義することによって，ランキング問題が分類問題に帰着できることは明らかである：$D'_k = \{(\mathbf{x}_{k,j} - \mathbf{x}_{k,i}, 1) | (\mathbf{x}_{k,i}, \mathbf{x}_{k,j}) \in D_k\}$．これは，$f_k(\mathbf{x}_{k,j} - \mathbf{x}_{k,i}) = f_k(\mathbf{x}_{k,j}) - f_k(\mathbf{x}_{k,i})$ であるためである．

3.1 節にて説明したエンティティランキング問題への適用は，訓練データとして $O'_{\preceq_k}$ 中のエンティティペアのベクトルを用いれば良い．すなわち，$D_k = \{(\mathbf{e}_i, \mathbf{e}_j) | (e_i, e_j) \in O'_{\preceq_k}\}$ とすることで，エンティティランキング問題へと適用できる．

3.4 文脈モデル

エンティティランキング問題に対する学習方法は既に述べたため，ここでは学習に用いられる文脈モデルについて説明する．文脈はあるラベル基準と特徴に関する文や文書の集合とすることができる．文脈を公開されているコーパスから得ることもできれば，代わりに Web をコーパスとして用いそこから得ることもできる．Web は巨大なコーパスであり，また，ラベル基準と特徴から適切なクエリを作ることができれば効率よくアクセスすることも可能である．そのため，本研究では，Web 検索結果から得られた文を用いて文脈を表現する方法について述べる．

ラベル基準 l_k と特徴 a_d が与えられた時，まず我々は l_k と a_d を AND 演算子で組み合わせたクエリを生成し，ある Web 検索エンジンへそのクエリを入力し上位 $N^{(c)}$ 件の検索結果を得る（実験においては $N^{(c)} = 500$ とした）．それから，検索結果のスニペットを文に分割し，ラベル基準 l_k と特徴 a_d の両方を含む文を抽出した．これらの文集合を以下では $S_{k,d}$ と表記する．

ラベル基準と特徴の文脈を表現する方法として 2 つの基本的な方法を提案する．1 つ目は TF-IDF 重み付け法によるベクトル表現であり，2 つ目は文の分散表現である [25]．両方法ともベクトルによって文脈を表現するが，先に述べたとおり，より複雑な表現方法も利用可能である．以下ではこれらの方法の詳細について簡潔に説明する．以下で用いられるいくつかの記号は，前の節での定義と異なった定義が与えられる．

語彙の集合を W とする．このとき，文集合 $S_{k,d}$ 中の各文 s_i はベクトル $\mathbf{v}_i$ によって表現され，$\mathbf{v}_i$ の j 次元目の値は以下のように定義される：$v_{i,j} = \mathrm{tf}(s_i, w_j) \log(n/\mathrm{df}(w_j))$．ただし，$\mathrm{tf}(s_i, w_j)$ は s_i 中に現れる語 $w_j \in W$ の頻度，$\mathrm{df}(w_j)$ は $S_{k,d} (k = 1, \ldots K; d = 1, \ldots, M)$ 中で w_j を含む文の数である．定数 n は $S_{k,d} (k = 1, \ldots K; d = 1, \ldots, M)$ 中の文の数を表す．

文脈 $c_{k,d}$ は文集合 $S_{k,d}$ のベクトルの平均によって表すことができる．しかし，もしあるラベル基準と特徴のペアに対して少数の文しか得られない場合，それらのラベル基準と特徴にはあまり関係がない可能性が高い．つまり，$w_{k,d} \sim 0$ である可能性が高い．これを実現するため，文集合 $S_{k,d}$ のベクトルの平均へ文の数に基づきペナルティを与える．そのため，文脈 $c_{k,d}$ は以下のように表現される：

$$\mathbf{c}_{k,d} = \frac{d(|S_{k,d}|)}{|S_{k,d}|} \sum_{s_i \in S_{k,d}} \mathbf{v}_i, \tag{18}$$

ここで $d(x)$ はペナルティ関数である．関数 $d(x)$ は $x = 0$ の

とき 0，x に対して単調増加し，$x \to \infty$ のとき 1 に収束することが望ましい．そのため，我々は $d(x) = \tanh(\alpha x)$ をペナルティ関数として用いた．ただし，α は関数の勾配を調整するパラメータである（実験では $\alpha = 1.0$ とした）．もしあるラベル基準 l_k と特徴 a_d に対してわずかな文しか発見できなかった場合，$d(|S_{k,d}|)$ は $\mathbf{c}_{k,d}$ のノルムを減少させるため，$w_{k,d}$ を 0 に近づけることができる．

もう一方の方法は文の分散表現を用いる方法である [25]．ある文集合が与えられた時，分散メモリモデルは文と前に出現する T 語によって条件付けられた語の確率の平均を最大にするような語と文の分散表現を得る．本論文ではこの手法の詳細については割愛する．分散メモリモデルの中でも，語と文の分散表現の平均によって語の確率を計算する手法を用いて，文集合 $S_{k,d}$ 中の各文 s_i に対して，分散表現 $\mathbf{v}_i$ を得ることができる．あとは，式 18 と全く同じ方法により文脈 $c_{k,d}$ を表現するベクトルを得られる．

上記で述べた 2 つの手法は異なる性質を持つ．TF-IDF 重み付け法によるベクトル表現は疎で，文脈に出現した語そのものを表現することができる．対照的に，文の分散表現は密で，文脈に出現した語そのものを保持することはできない．そのため，実験ではこれら 2 つの手法を比較する．

4. 実　　験

我々のタスクには公開されたデータセットが存在しないため，まずデータセットの作成の概略と統計情報について説明する．それからベースライン手法を含む実験設定について紹介し最後に実験結果を示す．

4.1 データセット

Web や雑誌から自動，または，手動によって，3 クラスのエンティティに対する様々なエンティティ順序を抽出した．これら 3 クラスは都道府県，国，カメラである．クラス選定は以下の 4 つの理由に基づいた：(1) 順序基準と特徴に関する Web ページの量，(2) エンティティ順序の種類数，(3) 数値属性の量，(4) 統計情報の多様性．

都道府県と国に関してはエンティティ列を Web ページ中のランキング（各エンティティが 1 位，2 位等に順位付けされているコンテンツ）から自動的に抽出した．カメラに関しては，国内で発売されている 10 のカメラ雑誌からエンティティ順序を手動で抽出した．このような雑誌には，各カメラに様々な観点から点数や順位が与えられているため，それらに基づいてエンティティ列を構築した．上記の結果として得られたエンティティ列はそれぞれエンティティペアに変換され，部分的に観測された順序集合 $O_{\preceq_k}$ を得た．要素数が 5 未満であった順序集合を除き，最終的に 158 のエンティティ順序を得た．

都道府県と国の数値属性は Web 上で公開されているものを自動的に抽出し，カメラについては価格.com(注1) のページよりスペック情報などを抽出し数値属性とした．数値属性を抽出後に，各属性ごとに $[0, 1]$ に収まるよう各属性の最大値と最小値によって正規化を行った．

表 1　各クラスの統計情報とエンティティ，エンティティ順序，および，属性の例．

	都道府県	国	カメラ
# Entities	47	138	149
# Orders	64	40	54
# Entities / Order	13.3	17.7	14.4
# Attributes	137	83	16
(# Entities / Order) / # Attributes	0.097	0.213	0.900
Entity examples	東京 京都	デンマーク アイスランド	EOS 5DS Nikon D3300
Attribute examples	人口 犯罪率	旅行者数 自殺者数	解像度 重さ
Order examples	魅力度 豊かさ	住みやすさ 幸福度	持ち運びやすさ 耐久性

表 1 に統計情報とエンティティ，エンティティ順序，および，属性の例を示す．多くのエンティティ順序に対して，我々はすべてのエンティティを含むようなランキングを見つけることができなかった．これは，それらのページが上位 1，3，5 件のみを紹介することが多かったためである．また，表中に示されている統計値の中で最も重要な値は属性数に対する 1 エンティティ順序あたりのエンティティ数の比率である（(# Entities / Order) / # Attributes）．この比率は全てのクラスで 1.0 以下であり，カメラクラスが最も高く，都道府県クラスが最も低くなっている．

4.2 実験設定

我々は実験におけるベースライン手法として，文脈を用いない既存のランキング学習手法を用いた：(1) **RankNet** [6]: ニューラルネットワークを用いクロスエントロピー損失を最適化するペアワイズランキング手法，(2) **RankBoost** [18]: AdaBoost [19] のペアワイズランキングへの応用，(3) **LinearFeature** [29]: 座標上昇法によって最適化された線形特徴モデル，(4) **LambdaMART** [33]: ランキング学習手法である LambdaRank [7] と決定木ブースティングモデルである MART [20] を組み合わせた手法，(5) **ListNet** [8]: ニューラルネットワークを用いたリストワイズランキング手法．我々は RankLib(注2) での実装を実験に用いた．なお，タスクが互いに関連しているという仮定をしなかったため，マルチタスク学習との比較は行わなかった．

構築したデータセットを用いた実験の手順を以下で述べる．各順序集合 $O_{\preceq_k}$ に対し，$O_{\preceq_k}$ に含まれるエンティティ集合 E を 1 対 1 に分割し，これらを E_{train} および E_{test} とした．訓練データは $O_{\text{train}} = \{(e_i, e_j) | (e_i, e_j) \in O_{\preceq_k} \wedge e_i \in E_{\text{train}} \wedge e_j \in E_{\text{train}}\}$ であり，テストデータは $O_{\text{test}} = O_{\preceq_k} - O_{\text{train}}$ とした．実験でのタスクは，O_{train} に基づいて順序モデルを学習し，各 $(e_i, e_j) \in O_{\text{test}}$ に対して e_i と e_j のどちらが上位に順序づけされるかを予測する問題とした．この問題において，精度は正しく予測できたエンティティペア数と定義された．我々はあるクラスにおけるエンティティ順序に対して，5 分割交差法を用い

(注1)：http://kakaku.com/

(注2)：https://sourceforge.net/p/lemur/wiki/RankLib/

表 2　各ランキング手法の精度 (± 標準誤差)．太字は各エンティティクラスでの最大の精度を表す．

	精度			
	都道府県	国	カメラ	総合
RankNet [6]	0.482 (0.023)	0.478 (0.025)	0.530 (0.030)	0.497 (0.015)
RankBoost [18]	0.513 (0.028)	0.636 (0.024)	0.552 (0.036)	0.557 (0.018)
LinearFeature [29]	0.566 (0.019)	0.670 (0.024)	0.614 (0.034)	0.609 (0.015)
LambdaMART [33]	0.614 (0.021)	0.659 (0.019)	0.697 (0.024)	0.654 (0.013)
ListNet [8]	0.559 (0.020)	0.518 (0.022)	0.504 (0.031)	0.530 (0.014)
CGL (TF-IDF, Linear)	**0.661** (0.017)	0.716 (0.022)	**0.823** (0.019)	**0.730** (0.012)
CGL (TF-IDF, RBF)	**0.661** (0.019)	0.725 (0.021)	0.799 (0.019)	0.724 (0.012)
CGL (Distributed, Linear)	0.646 (0.020)	0.701 (0.023)	0.798 (0.021)	0.712 (0.013)
CGL (Distributed, RBF)	**0.661** (0.018)	**0.731** (0.022)	0.804 (0.021)	0.728 (0.013)

て各手法の最適なパラメータを決定した．

文脈誘導型学習（**CGL**）では以下の設定を用いた．文脈モデルとして **TF-IDF** と **Distributed** ($L = 400$ の分散表現) を用いた．パラメータ c と C は上述の交差法を用いて決定された．式 16 中のカーネル κ として，線形カーネル (**Linear**) および RBF カーネル (**RBF**) を採用し比較した．

4.3　実験結果

表 2 に精度と標準誤差を示す．文脈誘導型学習（CGL）はどの設定においてもベースライン手法を上回る精度を達成している．文脈誘導型学習に基づいた手法の中で最も精度の高かった手法は CGL (TF-IDF, Linear) であり，次点は CGL (Distributed, RBF) であった．最も精度の高いベースライン手法である LambdaMart からの総合的な精度改善は 11.6%であった．ランダム化 Tukey HSD テスト [9][(注3)]（$\alpha = 0.01$）によれば，CGL (TF-IDF, Linear) と全ベースライン手法との差は統計的有意であり，文脈誘導型学習に基づいた手法の中では統計的有意差は認められなかった．

クラスごとに見ると，都道府県，国，カメラクラスにおいて，CGL (TF-IDF, Linear) は LambdaMART の精度をそれぞれ 8%，11%，18%改善した．我々は，事前に行った人工データでの実験と文脈モデルに用いられた 1 属性当たりの文数がそれぞれ 36.0，45.7，137 であったことから，文脈の質と量が文脈誘導型学習の効果を決定する主な要因ではないかと考えている．

我々は学習された関数内で用いられている属性の評価も行った．最も絶対値の高い重みを持つ 5 つの属性をエンティティ順序ごとにプールし，クラウドソーシングサービス Lancers[(注4)] のユーザに対して提示した．この実験では学習された属性がどの程度順序を説明できるのかを理解することを目的としている．ユーザは与えられた順序基準と属性の間に相関があるかどうかを，−2 から +2 までの 5 段階で評価した．我々は各順序基準と属性のペアに対して 5 名のユーザを割り当てた．最も精度の高かった CGL (TF-IDF, Linear) とベースライン中で唯一線形関数を用いる LinearFeature が評価に選ばれた．

(注3)：http://www.f.waseda.jp/tetsuya/tools.html
(注4)：http://www.lancers.jp/

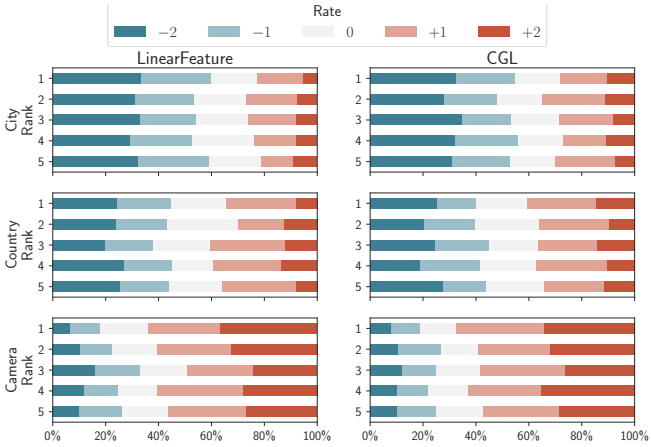

図 1　最も絶対値の高い重みを持つ 5 つの属性の評価値の分布．

表 3　文脈誘導型学習によって学習された線形関数の例．最も絶対値の高い重みを持つ 3 つの属性のみを示している．

クラス	学習された線型モデル							
都道府県	魅力度	=	+0.035	女性の平均寿命	−0.032	交通事故死者数	−0.031	人口／世帯数
都道府県	貯蓄額	=	−0.174	最高気温	+0.160	健康寿命	+0.148	旅館数
国	評判	=	+0.058	幸福度	−0.057	難民申請数	−0.045	自殺者数
国	平和	=	+0.170	穀物収穫量	+0.166	GDP 成長率	−0.126	自殺者数
カメラ	操作性	=	−0.240	重さ	−0.213	高さ	+0.133	シャッタースピード

図 1 に最も絶対値の高い重みを持つ 5 つの属性の評価値の分布を示す．都道府県，国，カメラに対して，文脈誘導型学習の平均評価値はそれぞれ，−0.455，−0.166，+0.581 であり，LinearFeature の平均評価値はそれぞれ，−0.560，−0.204，+0.516 であった．これらの評価値は手法の精度と高い相関がある．文脈誘導型学習はすべてのクラスにおいて LinearFeature よりも妥当な属性を発見できていたが，全てのクラスにおいてその差はわずかである．都道府県と国における平均評価値は 0 以下となっており，各属性の低い説明可能性を示唆している．これは単に各属性が順序基準と相関があったものの，順序基準に変化をもたらす要因として評価されなかったためではないかと考えている．文脈誘導型学習はベースラインよりも精度が高いモデルを学習できたものの，与えられた順序基準に対して高い説明可能性を持った属性を発見するのは未だ挑戦的な課題であると言える．

最後に，文脈誘導型学習によって学習された線形関数の例を表 3 に示す．多くの属性はエンティティ順序を説明可能で影響を与えているように見える．一方，いくつかの属性はエンティティ順序を説明するために適切でないように見えるが（例えば，“魅力度” に対する “世帯当たりの人口” や “貯蓄額” に対する “最高気温”），データセット中ではエンティティ順序と高い相関を示している．これらは主観的評価では妥当でないと判断されたものの，予測においては大きく貢献している属性の例である．

5. ま と め

本論文では，部分的に観測された順序を訓練データ，エンティティの数値属性を特徴として線形関数を学習することで，エンティティの順序を学習する問題に取り組んだ．この問題では，訓練データの数が非常に少ないため，我々は文脈誘導型学習という機械学習手法を提案した．文脈誘導型学習は，ラベル付けされたデータだけでなく，各特徴の重みを学習するためにラベル基準と特徴に関する文脈を利用する．実験では文脈誘導型学習によって既存のランキング学習手法を上回る高い精度が得られることが明らかとなった．

今後の課題としては，文脈誘導型学習の理論的解析，文脈誘導型学習の他の問題への応用（マルチラベル学習やゼロショット学習），より良い文脈モデルの検討，大規模データに対応するため効率化などを挙げる．

謝辞 本研究の一部は，文科省科研費 15H01718, 26700009, 16H02906, 16K16156, 25240050, 24680008 によるものです．ここに記して謝意を表します．

文 献

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[2] K. Balog, P. Serdyukov, and A. P. d. Vries. Overview of the TREC 2010 entity track. In *TREC*, 2010.

[3] K. Balog, P. Serdyukov, and A. P. d. Vries. Overview of the TREC 2011 entity track. In *TREC*, 2010.

[4] K. Balog, A. P. d. Vries, P. Serdyukov, P. Thomas, and T. Westerveld. Overview of the TREC 2009 entity track. In *TREC*, 2009.

[5] F. Bovolo, L. Bruzzone, and M. Marconcini. A novel context-sensitive svm for classification of remote sensing images. In *IGARSS*, pages 2498–2501, 2006.

[6] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005.

[7] C. J. Burges, R. Ragno, and Q. V. Le. Learning to rank with nonsmooth cost functions. In *NIPS*, pages 193–200, 2006.

[8] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, pages 129–136, 2007.

[9] B. A. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS*, 30(1):4, 2012.

[10] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[11] W. W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization. *ACM TOIS*, 17(2):141–173, 1999.

[12] H. Daumé III. Bayesian multitask learning with latent hierarchies. In *UAI*, pages 135–142, 2009.

[13] D. Davidov and A. Rappoport. Extraction and approximation of numerical attributes from the web. In *ACL*, pages 1308–1317, 2010.

[14] A. P. De Vries, A.-M. Vercoustre, J. A. Thom, N. Craswell, and M. Lalmas. Overview of the INEX 2007 entity ranking track. In *INEX*, pages 245–251, 2007.

[15] G. Demartini, A. P. de Vries, T. Iofciu, and J. Zhu. Overview of the INEX 2008 entity ranking track. In *INEX*, pages 243–252, 2008.

[16] G. Demartini, T. Iofciu, and A. P. De Vries. Overview of the INEX 2009 entity ranking track. In *INEX*, pages 254–264, 2009.

[17] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *KDD*, pages 109–117, 2004.

[18] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4:933–969, 2003.

[19] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1(55):119–139, 1997.

[20] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[21] T. Iwanari, N. Yoshinaga, N. Kaji, T. Nishina, M. Toyoda, and M. Kitsuregawa. Ordering concepts based on common attribute intensity. *IJCAI*, pages 3747–3753, 2016.

[22] L. Jacob, J.-p. Vert, and F. R. Bach. Clustered multi-task learning: A convex formulation. In *NIPS*, pages 745–752, 2009.

[23] C. Kang, D. Yin, R. Zhang, N. Torzec, J. He, and Y. Chang. Learning to rank related entities in web search. *Neurocomputing*, 166:309–318, 2015.

[24] A. Kumar and H. Daumé III. Learning task grouping and overlap in multi-task learning. In *ICML*, pages 1383–1390, 2012.

[25] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196, 2014.

[26] S.-I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller. Learning a meta-level prior for feature relevance from multiple related tasks. In *ICML*, pages 489–496, 2007.

[27] A. Madaan, A. Mittal, G. R. Mausam, G. Ramakrishnan, and S. Sarawagi. Numerical relation extraction with minimal supervision. In *AAAI*, pages 2764–2771, 2016.

[28] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE TAC*, 3(2):184–198, 2012.

[29] D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.

[30] Y. Qi and R. W. Picard. Context-sensitive bayesian classifiers and application to mouse pressure pattern classification. In *ICPR*, pages 448–451, 2002.

[31] H. Takamura and J. Tsujii. Estimating numerical attributes by bringing together fragmentary clues. In *HLT-NAACL*, pages 1305–1310, 2015.

[32] T. A. Tran, C. Niederée, N. Kanhabua, U. Gadiraju, and A. Anand. Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In *CIKM*, pages 1201–1210, 2015.

[33] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, 2010.

[34] K. Yu, V. Tresp, and A. Schwaighofer. Learning gaussian processes from multiple tasks. In *ICML*, pages 1012–1019, 2005.

[35] M. Zhou, H. Wang, and K. C.-C. Change. Learning to rank from distant supervision: Exploiting noisy redundancy for relational entity search. In *ICDE*, pages 829–840, 2013.