Time-Aware Information Retrieval in Social Networks

Miyanishi, Taiki

(Degree)
博士（工学）

(Date of Degree)
2014-03-25

(Date of Publication)
2016-03-25

(Resource Type)
doctoral thesis

(Report Number)
甲第6106号

(URL)
https://hdl.handle.net/20.500.14094/D1006106

（別紙様式 3）

論文内容の要旨

氏　　名　宮西　大樹

専　　攻　計算科学

論文題目（外国語の場合は，その和訳を併記すること。）

Time-Aware Information Retrieval in Social Networks

ソーシャルネットワーク上の時間情報を考慮した情報検索

指導教員　上原　邦昭

（注）2，000 字～4，000 字でまとめること。

World Wide Web 中のソーシャルネットワーク上で作成されるデータは，日々蓄積され，更新されていく．このような時々刻々と変化する大量のデータの中から個人が興味や関心を持つ重要な情報だけを入手するためには，ソーシャルネットワークの特性を考慮した情報検索手法が必要である．そこで，本稿ではソーシャルネットワークが持つリアルタイム性とネットワーク構造によって決まるユーザの重要度を考慮した情報検索の枠組みについて紹介する．

ソーシャルネットワークは，興味深い出来事が発生すると多くの人々が即座にその出来事について言及するリアルタイム性を有する．例えば，ある地域で地震が起こると，ほぼ同時にその地域付近に住む人々が身の回りの状況を報告・共有するために「地震 ○○市」，「地震 被害」，「○○市 避難所」などのキーワードを含む文書を作成する傾向がある．私の研究では，このリアルタイム性を利用することで，文書のタイムスタンプからユーザがいつの情報を欲しがっているかを推定し，過去と現在の情報を区別することにより，特定の時間に紐付いた情報の検索を行う．

しかし，特定の時間に関連する情報を検索するだけでは，重要でない文書も検索される可能性が高い．ソーシャルネットワーク中では，重要なユーザほど質の高い文書を作成する傾向があり，そのユーザの重要度は，ユーザ同士の関係が常に変化するため日々変化する．そこで，本稿では将来のネットワークを予測することで，ネットワーク構造によって決まるユーザの重要度推定方法について紹介する．以下に本稿の章構成を記載する．

1 章では，本研究の動機と貢献および本稿の概要について述べる．

2 章では，本研究の関連研究および本稿で用いる数理モデルの背景と評価方法について述べる．

3 章では，特定の時間に関連する情報を検索するためのクエリ拡張手法について紹介する．従来の検索は検索キーワードと内容的に一致する文書を検索することが目的であるため，特定の時間に関連する情報を検索することが苦手であった．そこで，検索キーワードがどの時間帯の情報を探しているかについて，その検索キーワードを多く含む文書のタイムスタンプの日毎の頻度から推定し，推定した時間帯に作成された文書中の単語を用いることで，特定の時間に関連する情報の検索を行う．ソーシャルネットワークの代表的なサービスである Twitter のデータ（Tweets2011 コーパス）を用いた実験から，本手法を用いることで従来の時間を考慮しない検索手法よりも検索精度を向上できることを示す．

4 章では，単語や 2 語以上の単語の組合せであるコンセプトを用いた疑似適合フィードバックによるクエリ拡張手法について紹介する．従来のクエリ拡張は，元のクエリを拡張するために単語を用いる．しかし，単語は意味的・時間 的曖昧性を持ちやすいため，元のクエリに意味的に関連のない単語を誤って関連すると判定する場合がある．そこで，単語よりも曖昧性の少ないコンセプトを用い，コンセプトの頻度を時間ごとに追跡することで，既存のコンセプトに基づく疑似適合フィードバックを拡張する．本手法を用いれば，単語よりも適合文書の検索に適した特

定の時間に頻繁に言及された重要なコンセプトを同定することができる．Tweets2011 コーパスを用いた実験により，提案するコンセプトに基づくクエリ拡張手法を用いることで，検索クエリに適合し，かつ情報量の豊富な文書を効果的に検索できることを示す．

５章では，情報検索において，疑似適合フィードバックにより検索精度が低下する問題の解決方法について紹介する．疑似適合フィードバックでは，初期検索の上位の検索結果は適合文書であり，この適合文書の中にユーザクエリの補強に役立つ単語が含まれていると仮定している．しかし，上位の検索結果の多くが非適合文書である場合，疑似適合フィードバックを用いると，元のクエリに関係のない単語が選ばれてしまう可能性がある．そこで，提案手法は上位の検索結果の中から適合文書を 1 つだけをユーザが選び，この文書をクエリ拡張に用いることで選んだ適合文書と類似した適合文書を上位の検索結果に集める．そして，再検索した上位の結果に対して疑似適合フィードバックを適用することで，疑似適合フィードバックによって検索精度が低下するクエリの数を抑えつつ，全体的な検索精度の向上させる．Tweets2011 コーパスを用いて提案手法である２段階の適合フィードバックの有効性を示す．

６章では，リンク予測の問題を解くことで，ソーシャルネットワーク中のユーザの順位を予測するモデルについて紹介する．従来では，ある時点のソーシャルネットワーク中のユーザをノード，ユーザ同士の関係をエッジとしたネットワークから，ネットワークの構造的な特徴を基に重要度や影響力の大きなユーザの同定を行ってきた．しかし，ユーザ同士の関係は年を追うごとに変化しており，ユーザの最新の重要度や影響力を把握するためには，現時点におけるユーザ間の関係を見るだけでは不十分である．そこで，本章では時間とともに変化するソーシャルネットワークを対象として，ネットワークの構造によって決定された各ノードの将来的な重要度・影響力（ネットワークの中心性）をリンク予測と RankBoost で将来のノードの順位を予測する．arXiv (hep-th) のデータセットから抽出したソーシャルネットワークを用いた実験により，本手法を用いることで将来的なノードの順位をより正しく予測できることを示す．

最後に，７章において本稿の結論と今後の展開について述べる．

（別紙１）　　論文審査の結果の要旨

氏名	宮西 大樹		
論文題目	Time-Aware Information Retrieval in Social Networks （ソーシャルネットワーク上の時間情報を考慮した情報検索）		
審査委員	区　分	職　名	氏　　名
	主　査	教授	羅志偉
	副　査	教授	有木康雄
	副　査	教授	上原邦昭
	副　査	准教授	関和広
	副　査		

要　　旨

World Wide Web 中のソーシャルネットワーク上で作成されるデータは，日々蓄積され，更新されていく．このような時々刻々と変化する大量のデータの中から個人が興味や関心を持つ重要な情報だけを入手するためには，ソーシャルネットワークの特性を考慮した情報検索手法が必要である．そこで本論文では，ソーシャルネットワークが持つリアルタイム性とネットワーク構造によって決まるユーザの重要度を考慮した情報検索の枠組みについて論じている．

ソーシャルネットワークは，興味深い出来事が発生すると多くの人々が即座にその出来事について言及するリアルタイム性を有することが知られている．例えば，ある地域で地震が起こると，ほぼ同時にその地域付近に住む人々が身の回りの状況を報告・共有するために「地震 ○○市」，「地震 被害」，「○○市 避難所」などのキーワードを含む文書を作成することが多い．本論文では，このリアルタイム性を利用することで，文書のタイムスタンプからユーザがいつの情報を欲しがっているかを推定し，過去と現在の情報を区別することにより，特定の時間に紐付いた情報の検索を行っている．しかし，特定の時間に関連する情報を検索するだけでは，重要でない文書も検索される可能性がある．ソーシャルネットワーク中では，重要なユーザほど質の高い文書を作成する傾向があり，そのユーザの重要度は，ユーザ同士の関係が常に変化するため日々変化する．そこで，本論文では将来のネットワークを予測することで，ネットワーク構造によって決まるユーザの重要度推定方法についても提案している．以下に本論文の章構成を記載する．

１章では，本研究の動機と貢献および本稿の概要について述べている．続く２章では，本研究の関連研究および本稿で用いる数理モデルの背景と評価方法についてまとめている．そして３章では，特定の時間に関連する情報を検索するためのクエリ拡張手法について提案している．従来の検索は検索キーワードと内容的に一致する文書を検索することが目的であるため，特定の時間に関連する情報を検索することが苦手であった．そこで本研究では，検索キーワードがどの時間帯の情報を探しているかについて，その検索キーワードを多く含む文書のタイムスタンプの日毎の頻度から推定し，推定した時間帯に作成された文書中の単語を用いることで，特定の時間に関連する情報を検索している．ソーシャルネットワークの代表的なサービスである Twitter のデータ（Tweets2011 コーパス）を用いた実験から，本手法を用いることで従来の時間を考慮しない検索手法よりも検索精度を向上できることが示されている．

４章では，単語や２語以上の単語の組合せであるコンセプトを用いた疑似適合フィードバックによるクエリ拡張手法について提案している．従来のクエリ拡張は，元のクエリを拡張するために単語を用いる．しかし，単語は意味的・時間的曖昧性を持ちやすいため，元のクエリに意味的に関連のない単語を誤って関連すると判定する場合がある．そこで，単語よりも曖昧性の少ないコンセプトを用い，コンセプトの頻度を時間ごとに追跡することで，既存のコンセプトに基づく疑似適合フィードバックを拡張する．本手法を用いれば，単語よりも適合文書の検索に適した特定の時間に頻繁に言及された重要なコンセプトを同定することができる．前述の Tweets2011 コーパスを用いた実験により，提案するコンセプトに基づくクエリ拡張手法を用いることで，検索クエリに適合し，かつ情報量の豊富な文書を効果的に検索できることが示されている．

氏名	宮西 大樹

５章では，情報検索において，疑似適合フィードバックにより検索精度が低下する問題の解決方法について議論している．疑似適合フィードバックでは，初期検索の上位の検索結果は適合文書であり，この適合文書の中にユーザクエリの補強に役立つ単語が含まれていると仮定している．しかし，上位の検索結果の多くが非適合文書である場合，疑似適合フィードバックを用いると，元のクエリに関係のない単語が選ばれてしまう可能性がある．そこで，本章の提案手法では上位の検索結果の中から適合文書を１つだけをユーザが選び，この文書をクエリ拡張に用いることで選んだ適合文書と類似した適合文書を上位の検索結果に集める．そして，再検索した上位の結果に対して疑似適合フィードバックを適用することで，疑似適合フィードバックによって検索精度が低下するクエリの数を抑えつつ，全体的な検索精度の向上を図っている．この手法についても，同じく Tweets2011 コーパスを用いてその有効性が示されている．

６章では，リンク予測の問題を解くことで，ソーシャルネットワーク中のユーザの順位を予測する枠組について提案している．従来では，ある時点のソーシャルネットワーク中のユーザをノード，ユーザ同士の関係をエッジとしたネットワークから，ネットワークの構造的な特徴を基に重要度や影響力の大きなユーザの同定を行ってきた．しかし，ユーザ同士の関係は年を追うごとに変化しており，ユーザの最新の重要度や影響力を把握するためには，現時点におけるユーザ間の関係を見るだけでは不十分である．そこで，本章では時間とともに変化するソーシャルネットワークを対象として，ネットワークの構造によって決定された各ノードの将来的な重要度・影響力（ネットワークの中心性）をリンク予測と RankBoost で予測している．arXiv (hep-th) のデータセットから抽出したソーシャルネットワークを用いた実験により，本手法を用いることで将来的なノードの順位をより正しく予測できることが示されている．

最後に，７章では本論文の結論と今後の研究の展開について述べられている．

本研究は，近年世界的に注目され，重要性が高まっているソーシャルネットワークサービスについて，リアルタイム性に着目して効果的な情報検索手法を研究したものであり，ソーシャルネットワークのノード（ユーザ）および情報（コンテンツ）の時間的特性について重要な知見を得たものとして価値ある成果の集積である．以上より，提出された論文はシステム情報学研究科学位論文評価基準を満たしており，学位申請者の宮西大樹は，博士（工学）の学位を得る資格があると認める．