潜在的意味を考慮したグラフに基づく複数文書要約

北島 理沙　　小林 一郎
お茶の水女子大学大学院人間文化創成科学研究科理学専攻
kitajima.risa@is.ocha.ac.jp　*koba@is.ocha.ac.jp*

概要　近年，情報技術の発展に伴い大量のテキストデータが蓄積されるようになり，その中からユーザが必要としている情報を選択することが必要となってきた．そして，情報を取捨選択するための一手法として，自動文書要約技術の必要性が高まっている．特に複数文書要約は，大量のデータの概要をユーザが捉えることが可能になるという点で，今後ますます重要となると考えられる．要約手法としては様々な手法が提案されている一方で，LexRank のようなグラフベースの要約手法の有用性が示されている．これは，文をノード，文間の類似度をエッジとしたグラフ表現において，固有ベクトル中心性の概念に基づいて文の重要度を計算する手法である．しかし，この手法が用いているのは文の単語頻度ベクトルのような表層的な情報のみであり，文のもつ潜在トピックは考慮していない．本研究では，潜在トピックを考慮したグラフを用いた複数文書要約手法を提案する．そして，DUC2004 を用いた実験を通して従来の手法である LexRank との比較を行い，潜在トピックがグラフベースの要約手法において有用であることを示す．

キーワード　複数文書要約，Latent Dirichlet Allocation，PageRank

1　はじめに

近年，情報技術の発展に伴って大量のテキストデータが蓄積されるようになり，適した情報を効率よく選択することが重要になってきている．そのため，人々が必要としている情報を選択するために自動要約の技術の必要性がますます高まっている．自動要約技術においては，様々な手法が提案されている一方で，文の関係のグラフ表現における固有ベクトル中心性の概念に基づいて文の重要度を計算する，グラフベースの文書要約技術が提案されており，その有用性が知られている．特に，LexRank [11] はリード手法 [7] や中心性に基づいた手法 [23] のようなベンチマーク手法として用いられる様々な手法よりも良い結果を示すことが知られている．この手法は文間の類似度を計算するのに表層情報に対するコサイン類似度を用いている．本研究では，潜在トピックに基づいた文の類似度グラフを用いる複数文書要約手法を提案し，DUC2004 [1]を用いた実験を通して LexRank よりも高い精度で複数文書要約を行う手法を提案する．

2　関連研究

自動要約技術としては，多くの手法がこれまでに提案されてきている一方で，文の類似度をグラフ表現したものを用いる手法が高い精度で文書要約を行えることが知られている [11, 20, 22]．Erkan ら [11] および Mihalcea ら [20] は，対象となる文書の概要をまとめた要約生成を行っており，前者では複数文書を，後者では単一文書を対象としている．加えて，Otterbacher ら [22] は，クエリに特化した要約生成を行っている．これらの研究は，リードベース手法 [7] や中心性を用いた手法 [23] などのベースライン手法よりも高い精度を示している．さらに，グラフベースの要約手法としては，PageRank アルゴリズム [8] の概念を適用した様々な手法が研究されている．PageRank アルゴリズムでは，グラフにおけるノードは Web ページを，エッジは Web ページ間のリンクを表し，グラフにおける Web ページの重要性を計算する[2]．一方で，自動要約のために用いられるグラフは類似度グラフと呼ばれ，ノードは対象文書内の文を，エッジは文間の類似度を表し，対象文書内の文の重要度を計算する．要約技術において PageRank アルゴリズムを適用している代表的かつ早期の研究としては，LexRank [11] がある．LexRank を応用した研究としては，Otterbacher ら [22] による研究，Zhang ら [29] による研究がある．特に，Zhang ら [29] は自動要約のためのグラフベースなサブトピック分割アルゴリズム（GSPSummary）を提案し，対象文書群の内部に隠れたトピックの論理的構造に基づいて LexRank にサブトピックモデルを導入している．近年の研究として，Agirre ら [1] はパーソナライズ化された PageRank アルゴリズム [17] を用いた教師なしの語彙曖昧性解消のための手法を提案し，Yan ら [28] は PageRank アルゴリズムに基づいてツイートの推薦をする手法を提案している．さらに，Badrinath ら [3] は，グラフにおけるダイバーシティを考慮するためにパーソナライズ化された PageRank に対して負の値を強化する手法を提案した．

他方で，要約生成に潜在トピック推定を適用した研究が多く提案されている [2, 10, 25, 9]．彼らは潜在トピッ

[1] http://www-nlpir.nist.gov/projects/duc/guidelines/2004.html

[2] PageRank アルゴリズムの詳細については 3 節にて述べる．

ク推定手法として Latent Dirichlet Allocation (LDA) を用い，自動要約において LDA が有用であることを示した．Arora ら [2] と Chang ら [10] は対象文書内のトピックを捉えるために LDA を用いた手法を提案し，Tang ら [25] と Celikyilmaz ら [9] はクエリに特化した要約作成のために LDA を用いた手法を提案した．本研究では，グラフを用いた要約生成に対して潜在トピックを適用した手法を提案する．

3 PageRank

PageRank とは，Brin ら [8] によって提案された，Web ページ間に存在するハイパーリンク関係を利用することでページの順位付けを行うアルゴリズムである．Web のハイパーリンク構造は，Web ページをノード，ページ間のハイパーリンクをエッジとした巨大な有向グラフとして表現され，このグラフに基づいて計算された PageRank スコアによって各 Web ページの順位付けが行われる．PageRank において中心となっている概念が，他の重要な Web ページからリンクが張られている Web ページは重要である，という考え方である．この概念に基づき，ある Web ページ P_i の PageRank である $r(Pi)$ は，そのページを指している他の全てのページのもつ PageRank の総和となり，式 (1) で表される．

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|} \tag{1}$$

ここで，B_{P_i} はページ P_i にリンクが張られているページの集合，$|P_j|$ はページ P_j から張られているリンクの個数を表す．つまり，ある Web ページの PageRank をそのページが指している他の全ての Web ページの PageRank として与えるとき，そのリンク数で割った値を割り振っていることになる．しかし，式 (1) を計算する際に $r(P_i)$ の値が未知であるため，この式を反復的に処理することで全ての Web ページの PageRank を求める．$k+1$ 回目の反復における Web ページ P_i の PageRank は，式 (2) で表される．

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|} \tag{2}$$

各 Web ページへの嗜好性を一様に捉える時，初期値として全ての Web ページに対して $\frac{1}{n}$ の PageRank を与え，式 (2) を反復的に計算していくことで，最終的に収束し，全ての Web ページの PageRank を求めることができる．なお，この反復計算には，べき乗法を用いる．べき乗法とは，行列の主固有値と主固有ベクトルを見つけるための反復法であり，マルコフ連鎖の定常ベクトルがマルコフ行列の左側主固有ベクトルであること，および，求めたい PageRank ベクトル[3]が Web ページ間のリンク関係を表した推移行列をもつマルコフ連鎖の定常ベクトルであることより，PageRank の計算に用いられる．

上記が PageRank の基本モデルであるが，実際にはこれに対して確率的調整および原始的調整を行っている [8]．これにより，これ以上リンクの張られていない PDF ファイルや画像ファイルなどのノードからある一定の確率で他のノードへ移動できること，および，べき乗法の繰り返しによって唯一の定常ベクトルを見つけられることが保証されている．この調整の結果，PageRank の計算は式 (3) で表される．ここで，n は対象となっている Web ページの総数，α は任意のページへ移動する確率を制御するためのパラメータであり，制動係数 (damping factor) と呼ばれる．

$$r_{k+1}(P_i) = \alpha \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|} + (1-\alpha)\frac{1}{n} \tag{3}$$

このようにして，重要な Web ページからリンクが張られている Web ページは重要であるという概念に基づき反復計算を行うことで各 Web ページに PageRank が割り振られ，その値に基づいて Web ページの重要度を示す順位付けが行われる．

4 LexRank

LexRank は，Erkan ら [11] により提案された，PageRank [8] の概念に基づいた複数文書要約手法である．要約手法には，文書の全体像をまとめる要約と，ある視点に特化した内容をまとめる要約とがあるが，LexRank は前者の要約を対象としている．LexRank は，対象文書内の文のグラフ表現における固有ベクトル中心性の概念に基づいて文の重要度を計算する手法である．これは，単に次数の多いノードを評価するだけでなく，次数の多いノードと隣接しているノードの重要度についても考慮し，その分に比例して対象ノードを評価することができる．この手法では，文間のコサイン類似度に基づいた連結性行列が文のグラフ表現の隣接行列として使われており，その隣接行列の第 1 固有ベクトルの成分を各ノードの中心性を表すスコアと考える．Erkan らは，類似度グラフを生成する際に，上で述べたように枝の重みを利用した重み付きグラフとして表す手法の他に，その枝の重みに対して閾値 t を用いて枝刈りを行い，重みなしグラフとして表す手法を提案している．前者の手法は Cont．LexRank，後者の手法は LexRank と呼ばれている．LexRank および Cont．LexRank は，実際には

[3] PageRank を要素としたベクトル．

上述の処理のみで要約文を生成するのではなく，Radev ら [24] の提案した要約システムである MEAD [4]の内部に組み込み，冗長性削減のための指標などと組み合わせることで要約文を生成することを前提としている．本研究では，提案手法を Cont. LexRank と比較して評価を行う．なぜなら，Cont. LexRank においては類似度グラフにおける枝刈りのためのパラメータを調節する必要がないために，提案手法との比較を行いやすいと考えられるからである．本稿では，Cont. LexRank を LexRank と呼ぶことにする．

5 提案手法

5.1 TopicRank

LexRank では文間の類似度として $tfidf$ 値を要素とする文ベクトルのコサイン類似度を用いているのに対して，文のもつトピック分布の類似度を文間の類似度として用いる手法を提案し，これを TopicRank と呼ぶことにする．LexRank は，文間の類似度を計算するために文のもつ $tf * idf$ ベクトルに対してコサイン類似度を適用しているが，提案手法では潜在的意味の観点から文間の類似度を計算するために各々の文に割り当てられたトピック分布を用いる．文内のトピック分布を推定するための手法として，Latent Dirichlet Allocation (LDA) [6] を用いる．LDA とは，一つの文書に対して複数のトピックが存在すると想定した確率的トピックモデルであり，それぞれのトピックがある確率を持って文書上に生起するという考えの下，そのトピックの確率分布を導きだす手法である．文書内の各トピックは単語の多項分布として表現され，各文はトピックの多項分布として表現される．表 1 は，文間のトピック分布の類似度を要素にもった接続行列の一例である．文間の類似度は，5.2 節で述べている指標によって計算される．表 1 の例では，総文数は 7 であり，$s0$ は対象文書群に含まれる文のうち 0 番目の文を表す．

表 1 文間のトピック分布の類似度

	s0	s1	s2	s3	s4	s5	s6
s0	1.00	0.03	0.02	0.01	0.07	0.45	0.08
s1	0.03	1.00	0.37	0.23	0.02	0.30	0.24
s2	0.02	0.37	1.00	0.09	0.12	0.18	0.21
s3	0.01	0.23	0.09	1.00	0.14	0.03	0.19
s4	0.07	0.02	0.12	0.14	1.00	0.16	0.11
s5	0.45	0.30	0.18	0.03	0.16	1.00	0.27
s6	0.08	0.24	0.21	0.19	0.11	0.27	1.00

類似度グラフは文間の類似度を要素とした接続行列に基づいて生成される．例えば，図 1 は表 1 の類似度グラフを示しており，各ノードは文を表し，各エッジの重みは文間の類似度を表している．ここで，類似度グラフは文間の類似度を重みとした重みつきグラフとして示されている．

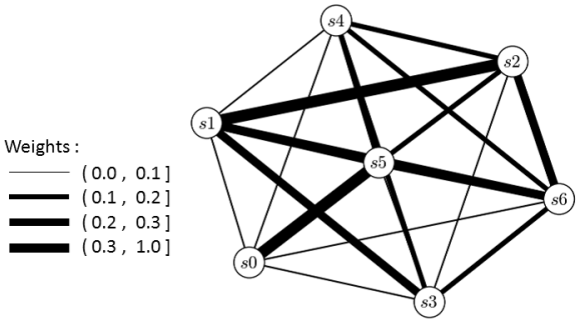

図 1 類似度グラフ

次に，生成された類似度グラフに対して，固有ベクトル中心性に基づいた各文の重要度を計算する．文 u の重要度は，Erkan ら [11] の手法を参考にして，式 (4) で求められる．ここで，N は対象としている文書群の総文数，$adj[u]$ は文 u の隣接ノード集合，d はある一定の割合で非隣接ノードとの類似度を考慮するための制動係数 (damping factor) である [8]．制動係数 d の値は，Brin ら [8] の結果を参考に $d = 0.85$ とした．類似度 $sim(u, v)$ の計算については，5.2 節に示す．

$$p(u) = d \sum_{v \in adj[u]} \frac{sim(u,v)}{\sum_{z \in adj[v]} sim(z,v)} p(u) + \frac{1-d}{N} \quad (4)$$

次に，重要度を要素とした行列に対してべき乗法を用いて第 1 固有ベクトルを計算する．これにより，中心性の高い文と類似していることがその文の重要度を高める，という概念に基づいた文の重要度を求めることができる．最後に，計算された重要度に基づいて文をランク付けし，上位から文を選択していくことで要約文が生成される．

5.2 文間類似度

文間の類似度として，LexRank では表層的な類似度のみを用いている一方で，提案手法である TopicRank では、例えば単語の出現頻度のような表層的な情報だけでなく，文書のもつ潜在的な情報によっても要約が生成されるように，対象文書群の表層的な類似度と潜在的な類似度の両方を用いる．式 (5) は，TopicRank の枠組みにおいて定義されている文 S と文 T の間の類似度を示している．P と Q は，それぞれ文 S と文 T のもつトピック分布である．式 (6) は，トピック分布に基づいた類似度を示している．文の重要度は，式 (5) を用いて

[4]http://www.summarization.com

式 (4) によって計算される．トピック分布の類似度指標としては，LDA において高い精度を示すことが知られている [15]Jensen-Shannon ダイバージェンスを用いる．実際には，コサイン類似度，Jensen-Shannon ダイバージェンス，そして Hellinger 距離の 3 つの指標を比較することによって，どの指標が提案手法に適しているかを考察するための予備実験を行った．そして，その中で Jensen-Shannon ダイバージェンスが提案手法においてトピック分布の類似度を計算するのに適していることが分かった．

$$sim(S,T) = \alpha * sim(P,Q) + (1-\alpha) * cosine(tfidf(S), tfidf(T)) \quad (5)$$

$$sim(P,Q) = 1 - D_{JS}(P,Q) \quad (6)$$

5.3 冗長性削減

重要文を抽出する際，TopicRank によって高い重要度をもった文のみを抽出していくと，冗長性のある要約文が生成される可能性がある．この問題を避けるために，提案手法では MMR-MD(Maximal Marginal Relevance-Multi Documents) [12] を応用した．この指標は，新しく抽出された文と既に抽出された文との間の類似度に対応するペナルティ値を与えることにより，類似した文を抽出することを防ぎ，クエリに特化した要約においてしばしば使用されている．これは，式 (7) によって定義される．ここで，Sim_1 は対象文書内の文とクエリとの間の類似度を表わし，Sim_2 は対象文書内の文と，生成した要約の一部として既に抽出された文との間の類似度を表わしている．2 つの類似度は，対象文書内の文とクエリの *tf*idf* ベクトルのコサイン類似度に基づいて計算されることが多い．この指標を用いると，与えられたクエリに類似した内容をもち，かつ，既に抽出された文と類似していない文を抽出することができる．

$$MMR\text{-}MD \equiv argmax_{C_i \in R \setminus S}[\lambda Sim_1(C_i, Q) - (1-\lambda) max_{C_j \in S} Sim_2(C_i, C_j)] \quad (7)$$

C_i ： 対象文書群内の文
Q ： クエリ
R ： クエリ Q によって検索された文集合
S ： 既に抽出された R 内の文集合
λ ： 重みパラメータ

本研究では，既に選択された文と類似した文を選択する機会を減らしつつ，TopicRank アルゴリズムによって計算された重要度をもつ文を選択することを目的とする．これを考慮し，MMR-MD において，潜在的な意味の観点において文書群における文の重要度を推定する Sim_1 として TopicRank スコアを，既に選択された文と抽出する文との間の類似度を計算する Sim_2 として文の *tf*idf* ベクトルのコサイン類似度を適用した．重みパラメータ λ については，後述する実験を通して適した値を考察する．本稿では，以後 MMR-MD を応用した TopicRank を TopicRank (MMR) と呼ぶことにする．

6 実験

6.1 実験設定

対象データには，DUC2004 の Task2 で使われた文書データを用いた．約 10 件の新聞記事からなる文書群が 50 セット用意されており，それらを用いて複数文書要約を行う．ここでは，LexRank と TopicRank を用いた結果を比較する．評価指標としては，それぞれの手法によって生成された要約に対して ROUGE [19] を適用する．特に，人間の評価と相関していることが示されている，ROUGE-1 値を用いる [19]．また，ストップワードを含めた値とストップワードを除いた値を求めることにし，前者を with，後者を without として示す．本実験においては，まず文間の類似度を計算するための式 (5) における重みパラメータ α の適切な値について考察する．そして，冗長性を削減するための式 (7) における重みパラメータ λ の適切な値についても考察する．その後，提案手法である TopicRank および TopicRank(MMR) を，従来の手法である LexRank と比較する．なお，本実験においてトピック数は 20 とする．LDA において潜在トピックの推定手法としては，ギブスサンプリングを用い，その反復回数は 200 とする．

6.2 実験結果

図 2 に，TopicRank における重みパラメータ α の変化に伴う ROUGE-1 値の変化を示す．図 2 によると，パラメータ α の値が増加するのにつれて ROUGE-1 値は増加しており，$\alpha = 1.0$ のときに ROUGE-1 値は最も高い値となっている．この結果から，TopicRank においては文間の表層的な類似度と潜在的な類似度の両方を用いるよりも，潜在的な類似度のみを用いる場合の方が精度の高い要約を生成できることが分かる．

図 3 に，冗長性削減を考慮した TopicRank における重みパラメータ λ の変化に伴う ROUGE-1 値の変化を示す．図 3 によると，$\lambda = 0.5$ のときに ROUGE-1 値は最も高い値となっている．この結果から，TopicRank における重みパラメータ λ の値は 0.5 と設定することに

する．

表 2 に，LexRank，TopicRank および TopicRank(MMR) 間での ROUGE-1 値の比較を示す．TopicRank および TopicRank(MMR) の場合の ROUGE-1 値は，LexRank の場合と比べて高くなっている．この結果より，類似度グラフのノードとして単語の頻度のような表層的な情報を用いるよりも，トピック分布のような潜在的な情報を用いる方が，要約生成において高い精度を示すことが分かる．さらに，TopicRank(MMR) は TopicRank よりも高い ROUGE-1 値を示しており，このことから，文のグラフ表現における文の重要度だけでなく，生成した要約の一部として既に抽出された文との類似度も考慮した MMR に基づいた提案手法である TopicRank(MMR) は，要約生成において冗長性削減のための手法として結果が向上していることが分かる．

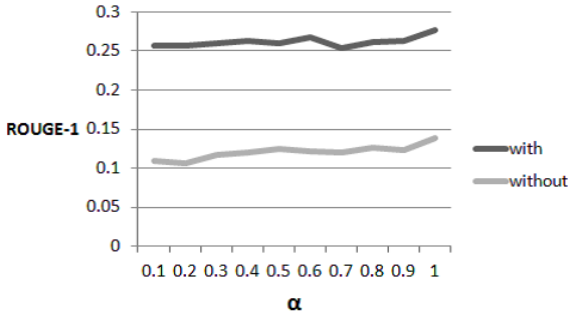

図 2 TopicRank における ROUGE-1 値の変化

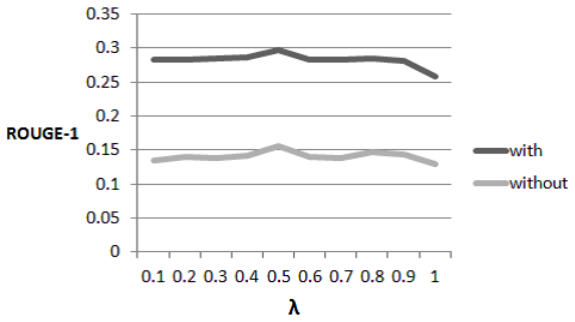

図 3 TopicRank(MMR) における ROUGE-1 値の変化．

表 2 ROUGE-1 値の比較

method	with	without
LexRank	0.222	0.035
TopicRank	0.276	0.139
TopicRank(MMR)	0.298	0.155

6.3 考察

実験結果より，グラフのアルゴリズムを用いた要約生成に対して，類似度グラフにおける類似度としてトピック分布を用いることが有用であることが分かった．潜在的な特徴に基づいた類似度と表層的な特徴に基づいた情報の適切な割合について考察した結果として，文のもつトピック分布のみを類似度グラフの表現に用いたほうが，潜在的な特徴と表層的な特徴の両方を用いるよりも良いことが分かった．さらに，TopicRank(MMR) の場合の実験結果より，類似度グラフを生成する際にはトピック分布のような潜在的な特徴を用いる方が良い一方で，冗長性を削減するために用いる特徴としては，単語の頻度のような表層的な情報を用いた方が，良い要約を生成できることが分かった．

7 おわりに

本研究では，複数文書要約手法である TopicRank を提案し，重要文を抽出する際に PageRank アルゴリズムを応用することで要約文生成を行った．この手法では，文を類似度グラフと呼ばれる文のグラフ表現におけるノードとして表し，文間の類似度をエッジとして表わした．他の多くの重要文と類似した文が重要文であるとみなされており，PageRank アルゴリズムによってグラフ内の文の中心性を計算することにより要約の一部として抽出される．特に，従来の単語の頻度のような表層的な特徴に基づいた類似度を計算する要約手法と異なる点として，文間の類似度を計算する際に文のもつトピック分布といった潜在的な情報を導入した．また，要約文として文を選択していく際に，文の表層的な特徴に基づいて冗長性を削減するために，MMR-MD に基づいた指標を提案した．最後に，DUC2004 を通して提案手法による要約文生成について考察するための実験を行い，ROUGE-1 値によって提案手法の評価を行った．実験結果から，要約文生成において LexRank[11, 29] よりも提案手法の方が高い精度を示すことを確認し，特に，提案した冗長性削減のための指標を用いた手法である TopicRank(MMR) においてはさらに高い精度で要約文生成を行えることが分かった．結果として，文を表わすノード間の類似度を計算するために潜在的な特徴を用いて類似度グラフを生成することは重要であり，生成された要約文の冗長性を削減するためには，*tf*idf* ベクトルのコサイン類似度のような表層的な特徴を用いた計算を行うことが重要であることが分かった．

今後の課題としては，潜在トピックの数や類似度グラフにおける枝刈りの仕方によって，どのように要約文の精度に影響が与えられるのかを考察することがあげられる．また，各文がどの文書から抽出されているのかに着

目し，文書-文間の関連性を考慮したモデルを提案手法に導入していきたいと考える．

参考文献

[1] Eneko Agirre and Aitor Soroa. : Personalizing PageRank for Word Sense Disambiguation, Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics, 2009.

[2] Rachit Arora and Balaraman Ravindran. : Latent Dirichlet Allocation based on Multi-document Summarization, Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data, 2008.

[3] Rama Badrinath and C. E. Veni Madhavan. : Diversity in Ranking using Negative Reinforcement. Mining Data Semantics in Information Networks (MDS) workshop, 2012.

[4] Harendra Bhandari, Masashi Shimbo, Takahiko Ito and Yuji Matsumoto. : Generic Text Summarization Using Probabilistic Latent Semantic Indexing, Proceedings of the 3rd International Joint Conference on Natural Language Processing, pp. 133–140, 2008.

[5] Qin Bing, Lin Ting, Zhang Yu and Li Sheng. : Research on Multi-Document Summarization Based on Latent Semantic Indexing, Journal of Harbin Institute of Technology, pp. 91–94, 2005.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, : Latent Dirichlet Allocation, Journal of Machine Learning Research, pp. 993–1022, 2003.

[7] Ronald Brandow, Karl Mitze and Lisa F. Rau, : Automatic Condensation of Electronic Publications by Sentence Selections, Information Processing and Management: an International Journal – Special issue: summarizing text, pp. 675–685, 1995.

[8] Sergey Brin and Lawrence Page. : The Anatony of a Large-scale Hypertextual Web Search Engine, Computer Networks and ISDN Systems, pp. 107–117, 1998.

[9] Asli Celikyilmaz and Dilek Hakkani-Tur, : Discovery of Topically Coherent Sentences for Extractive Summarization, Proceedings of the 49th Annual Meeting of the Association for Computatinoal Linguistics: Human Language Technologies, pp. 491–499, 2011.

[10] Ying-Lang Chang and Jen-Tzung Chein, : Latent Dirichlet Learning for Document Summarization, Proceeding of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1689–1692, 2009.

[11] Gunes Erkan and Dragomir R. Radev, : LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, Journal of Artificial Intelligence Research, pp. 457–479, 2004.

[12] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz, : Multi-document Summarization by Sentence Extraction, Proceedings of the 2000 NAALP-ANLP Workshop on Automatic Summarization, pp. 40–48, 2000.

[13] Thomas L. Griffiths and Mark Steyvers, : Finding Scientific Topics, Proceedings of The National Academy of Sciences, pp. 5228–5235, 2004.

[14] Aria Haghighi and Lucy Vanderwende, : Exploring Content Models for Multi-Document Summarization, Human Language Technologies: The 2009 Anual Conference of the North American Chapter of the ACL, pp. 362–370, 2009.

[15] Leonhard Hennig, : Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis, International Conference RANLP 2009-Borovers, pp. 144–149, 2009.

[16] Thomas Hofmann, : Probabilistic Latent Semantic Indexing, Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57, 2009.

[17] Glen Jeh and Jennifer Widom, : Scaling Personalized Web Search, Proceedings of the Twelfth International World Web Conference, 2002.

[18] Jianhua Lin, : Divergence Measures based on the Shannon Entropy, IEEE Transactions on Information Theory, pp. 145–151, 2002.

[19] Chin Y. Lin, : ROUGE: a Package for Automatic Evaluation of Summaries, Proceedings of the Workshop on Text Summarization Branches Out, pp. 74–81, 2004.

[20] Rada Mihalcea and Paul Tarau, : TextRank: Bringing Order into Texts, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 401–411, 2004.

[21] Ani Nenkova and Lucy Vanderwende, : The Impact of Frequency on Summarization, Technical report, Microsoft Research, 2005.

[22] Jahna Otterbacher, Gunes Erkan and Dragomir R. Radev, : Using Random Walks for Question-focused Sentence Retrieval, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 915–922, 2005.

[23] Dragomir R. Radev, Hongyan Jing and Malgorzata Budzikowska, : Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies, ANLP/NAACL Workshop on Summarization, 2000.

[24] Dragomir R. Radev, Sasha B. Goldensohn and Zhu Zhang, : Experiments in Single and Multi-document Summarization using MEAD, First Document Understanding Conference New Orleans, 2001.

[25] Jie Tang, Limin Yao and Dewei Chen, : Multi-topic Based Query-Oriented Summarization, Proceedings of SIAM International Conference on Data Mining, pp. 1147–1158, 2009.

[26] Yee W. Teh, David Newman and Max Welling. : A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation, Proceedings of the Advance in Neural Information Processing Systems Conference, 19, pp. 1353–1360, 2006.

[27] Xiaojun Wan and Jianwu Yang. : Improved Affinity Graph based Multi-document Summarization, Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, 2006.

[28] Rui Yan, Mirella Lapata and Xiaoming Li, : Tweet Recommendation with Graph Co-Ranking, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 516–525, 2012.

[29] Jin Zhang, Xueqi Cheng and Hongbo Xu, : GSP-Summary: A Graph-Based Sub-topic Partition Algorithm for Summarization, Information Retrieval Technology, vol. 4993, pp. 321–334, 2008.