KeyGraph による主張点の極性評価
－LDA の潜在トピックを用いて－

輪島 幸治[†1]　古川 利博[†2]　嶋田 茂[†1]

† 産業技術大学院大学 〒140-0011 東京都品川区東大井 1-10-40
‡ 東京理科大学 〒162-0825 東京都新宿区神楽坂 1-3

あらまし　近年，企業に対する質問文書等を早急に察知し，対応することで企業に対する信頼性を向上させることが求められている．ここで，ソーシャルメディア上の投稿から書き手の感情を定量化し，分析する手法が実現出来れば，その投稿に含まれるネガティブな感情の度合いから対応する優先順位付けを行うことが期待出来る．KeyGraph では，高頻度な単語群から語彙的結束性を用いて文章の概念を抽出し，その概念と共起する単語は文書において書き手の主張を表す重要な語としている．主張語を話題の絞込語として用いることで，書き手の感情を定量化することが期待出来る．しかし文章では，必ずしも高頻度な単語群に文章の概念が含まれているとは限らない．そこで，本研究では質問文書に対して LDA を適用し， 潜在的な意味に分かれた話題を抽出し，これと共起する単語を書き手の主張を表す語とする．この語を用いることで文書の優先付けを行う方法を提案する．

キーワード　KeyGraph，評価極性，ヘルプデスク，ソーシャルメディア，LDA

1. はじめに

近年，ソーシャルメディア上のデータを対象とした，データ解析分野の研究が盛んに行われている．その中でも着目を浴びつつある分野がテキストに含まれる感情分析である．ソーシャルメディアの書き手の感情は，データを解析する上で非常に重要である．テキストから，文字以外に書き手の感情を定量化することで，製品やサービスが時系列においてどのように顧客に評価されているかに関して分析することが可能となる．

2. 研究目的

本研究では，ウェブ上のサポートコミュニティと呼ばれる製品に関しての質問文書を対象に，ネガティブ度合いが高い文書を順位付けすることを目的としている．我々は文書に潜在的に含まれているネガティブさを深刻度と定義した[1]．感情に基づいた質問文書の順位付けが行えることで，企業に対する不満や製品クレームなどを早急に察知し，対応することが期待できると期待される．

3. 極性評価

極性評価とは，単語が持つポジティブさあるいはネガティブさの度合いを定量化した評価極性値を基に，文書のネガティブさやポジティブさを評価する手法である．関連研究としては，商品のレビューサイトの対象とした Turney らの研究[2][3]が代表的である．Turney らは文書に出現する単語の評価極性値の平均値(以後，平均評価極性値）を文書の評価に用いている．

極性値の算出には，極性値の妥当性確保や算出のための大規模な文書集合必要な点から，本研究では高村ら[4]が提案している単語感情極性対応表の極性値を評価極性値として用いる．

3.1. 極性評価の課題

これまでに極性評価を用いて質問文書の順位付けを検討した．その結果，文書内で否定的な単語を使用している質問文書に限っては，評価可能であった．

だが単語の意味としては肯定的だが，文書としては批判的な文書の場合などを適切に順位付けすることが困難であるという問題がある．

また文書内の話題に着目した評価を検討し，後述するトピックモデルを採用したトピックの極性評価を検討し，話題ごとに極性評価するアプローチや，話題と共起する単語から話題を極性評価するアプローチを検討した．だが文書を適切に順位づけするまでには至らなかった．

これは，話題そのものはポジティブとネガティブが混在しているニュートラルな場合が多く，話題ごとに一意的ではないことが考えられる．

そこで我々は，文書内に含まれる主張語に着目した．文書に含まれる主張に着目することで，文書が話題に対して，どのように主張していることが明らかになり，話題に対する主張ごとにネガティブ度合いの高い文書ごとに優先付けを行うことが可能であると期待できる．文書の主張語を抽出する手法として，KeyGraph を採用した．

4. KeyGraph

KeyGraph[5]とは，文書は著者の考えを主張するために書かれるという仮説を基に，検索対象の各文書から，文書の話題と主張を抽出する手法である．KeyGraph では文書中に出現する単語の出現頻度と共起関係から，文書中での内容展開を表現している単語のクラスタを作成し，文書中の話題とする．そのクラスタより筆者の主張点を抽出する．

筆者の主張を持つ単語を得る事で，ユーザは検索したい内に近い分野の文書を探すことができる．これは一般的な頻度解析等の統計的解析と違い，高頻度な単語だけでなく，低頻度な単語にも意味があると考えていることに起因しているためである．KeyGraph は以下 2 つの考え方に基づいて構成されている．

①文書中で繰り返し出現する頻度の高い単語は，その文書全体の内容展開の基本となる概念，文書の「土台」となる．

②文書中では，①の土台に基づいて，文書に筋道が与えられる．この筋道に支えられているのが，文書中で筆者が最も伝えたい主張である．この主張こそが文書のキーワードである．

KeyGraph の構成手順に関して説明する．

・KeyGraph の構成手順

Step1: 土台の抽出

文書を構成する単語の出現頻度を算出し，頻度の高い単語を一定数 M 個を取り出す．

Step2: 共起グラフの作成

頻度上位 M 個の単語同士の共起度 Co を測り，共起度の高い順にソートし，M-1 番目までに単語同士に枝を張る．共起度 Co に関しては Jaccard 係数を用いる．

Step3: 土台の抽出

共起グラフ中の対になるノード w_i，w_j を結ぶ枝に対して，この枝を切り離しても他の枝を遷移する事で w_i から w_j へ到達出来る場合は，そのまま枝を残す．到達出来ない場合は，この枝を切断する．単結合のパスは切除する．

Step4: 橋の抽出

元の文書集合の全ての単語（名詞、副詞、形容詞）と Step3 までで抽出した土台を構成する単語との共起度 C1(w，土台_i)を構成しているそれぞれの語を求める．

Step5: KeyGraph の抽出

KeyGraph では，文書集合を構成する単語 w（名詞、副詞、形容詞）が土台に支えられる力を Key(w)と表す．Key(w)は 0 ～ 1 までの実数である．ある語に対して Step4 で出した共起度 C1＞0 となる土台が 2 つ以上ある場合は以下の通りとする．

$$f(w,\ g)=\sum_{s\in D}|w|\_s|g_i-w|\_s$$
$$F(g)=\sum_{s\in D}\sum_{w\in s}|g_i-w|\_s$$

ここで，s と w をそれぞれ文と語を指す添え字，|g|_s を土台 g に含まれる語の s 中の出現回数として

$$|g-w|\_s=\min\begin{cases}|g|_s-|w|_s & \text{if } w\in g\\ |g|_s & \text{if } w\in g\end{cases}$$ とする．

すなわち，f(w，g)は語 w と土台 g 中の語の共起度である．Key は次の様に定義する．

$$Key(w)=\{1-\prod_g^{bases}(1-\frac{f(w,\ g\_i)}{F(g\_i)})\}$$

それ以外の場合，つまり共起度 C1>0 となる土台が 1 以下の場合に Key=0 とする．ここで，F は正規化係数である．F の値の最大値が 1 を超える場合，F は f の最大値とする．それ以外は 1 に設定する．

Step6: KeyGraph の抽出

Key 値が上位 R 個を筆者の主張を表す語とする．

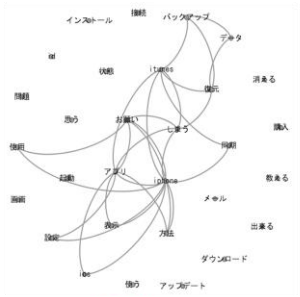
4.1. KeyGraph の課題

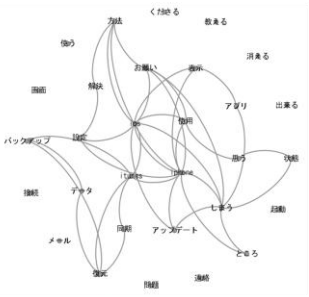
KeyGraph では，構成手順 Step2・Step3 にあるとおり，抽出された高頻度の単語群とそれらの共起から話題のクラスタを抽出する．しかしながら高頻度の単語群のみでは，文書群の話題を適切にクラスタ分けをすることが難しい場合がある．

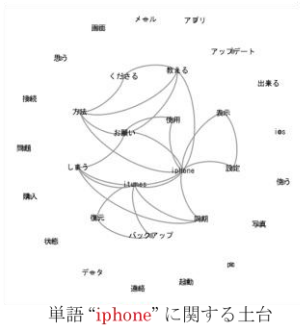
Apple サポートコミュニティのデータに対して，KeyGraph による土台抽出プロセスを実施した結果を次ページ図 1 に示す．図の"~に関する"の定義だが，対象となる文書集合から，キーワードで絞りこんだものを指す．例えば，"アプリに関する土台"とは，文書集合に対して，"アプリ"という単語を含む文書を絞り込んだものに対して，土台抽出プロセスを実施したものになる．

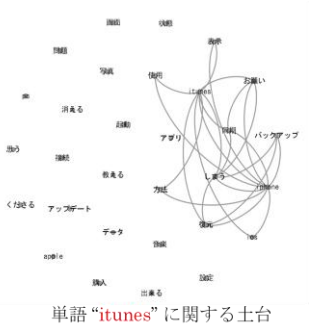
図 1 の結果からわかるように Keygraph による単結合のパスのプロセスを経た結果，1 つの大きな土台が形成される形となってしまうことが分かった．

そのため頻度や共起度だけでなく，適切なクラスタ分けのために，潜在的な意味に基づいて話題を分割する必要がある．そこで我々はこの土台抽出プロセスに対して，トピックモデルによる拡張を検討した．


単語 “アプリ” に関する土台


単語 “ios” に関する土台


単語 “iphone” に関する土台


単語 “itunes” に関する土台

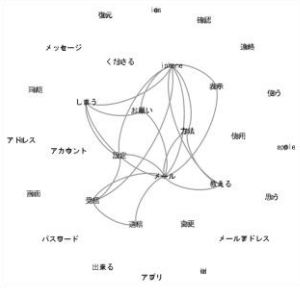

単語 “メール” に関する土台

図 1 Apple サポートコミュニティに対する土台抽出

5. トピックモデル

トピックとは，話題や分野など文書に含まれる特徴である．トピック抽出のための手法を一般にトピックモデルという．

トピックモデルには，文書ごとに単一のトピックを選ぶユニトピックモデルと文書に複数のトピックが割り当て可能なマルチトピックモデルがある．本研究では潜在的な情報からトピックを抽出することが可能な，マルチトピックモデルの一つである潜在的ディリクレ配分法 [5](Latent Dirichlet Allocation，以後 LDA)によるトピック抽出手法を採用する．

5.1. 潜在的ディリクレ配分法

LDA とは，文書中において単語は潜在的なトピックによって出現するという考えに基づいたマルチトピックの一つである．文章を構成するトピックがある確率を持って文書上に生起するという仮定の下，文章を構成する単語に対して，生起しやすいトピックを割り当てる事で，トピックの確率分布を導き出す手法である．

LDA は，自然言語処理の分野においては，文書分類に用いることが出来ることが知られている[4]．LDA による文書の生成過程は次のようになる．

・LDA の構成手順

Step1: Dirichlet 分布に従いトピックごとに生成する単語の確率分布を生成．

$k = 1, \cdots, K$

$\phi_k \sim Dir(\beta)$

以下，文書 d ごとに繰り返す（$d = 1, \cdots, M$）

Step2：ポアソン分布に従い文書 d の単語数 N を生成

$N \sim Poisson(\xi)$

Step3: Dirichlet 分布に従い各トピックを生成するトピックの確率分布（混合比）を生成．

$\theta_d \sim Dir(\alpha)$

Step4: Step2 で求めた N 個のそれぞれの単語 w_n に対し，以下を行う．

(1)Step3 で求めたトピックの確率分布に従い，トピック Z_{dn} を生成．

$Z_{dn} \sim Multinomial(\theta_d)$

(2)トピック Z_{dn} に対する単語の確率分布に従い単語 w_{dn} を生成．

よって LDA によるコーパス D の生成確率は以下の式で定義することができる．

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta|\alpha) \left(\prod_{n=1}^{N} p(Z_{dn}|\theta) p(w_{dn}|Z_{dn}, \beta) \right) d\theta_d$$

ここで，α，βはハイパーパラメータを表し，ユーザが文書集合に適した値を推定し入力する．

パラメータの推定する代表的な手法には，EM アルゴリズムやギブスサンプリングが挙げられるが本研究ではギブスサンプリングを採用している．

6. 提案手法

提案手法では，まず抽出フェーズとして，KeyGraph の土台として LDA を用いたトピックを土台として用いて，主張語を抽出する．次に割り当てフェーズとして，各文書に対して LDA のトピックを割り当てる．その後，文書内の主張語の有無から，各文書に主張語を割り当てる．評価フェーズにとして，文書に割り当てられた話題と主張語から，文書の順位付けを行う．順位付けの基準には，話題と主張語が割り当てられた文書群に対する平均評価極性値を用いる．

7. 実験

提案手法の有用性を検証するため，計算機によるシミュレーション実験を実施した．実験内容及び実験に使用したシミュレーション環境，解析対象，データクレンジングに関して記載する．

7.1. 実験内容

評価極性値の分布を検証するため，予備実験として単語感情極性対応表の極性分布，解析対象のデータの平均評価極性値，比較対象のデータの平均評価極性値に関して，極性分布を確認した．

その後，本実験として解析対象のデータに対して潜在トピックを抽出し，各文書に割り当て，KeyGraph による主張語の抽出を行った．

7.2. シミュレーション環境

シミュレーションに使用した OS は CentOS release 6.5 を使用し，実装には Python 2.6.6 を用いた．LDA アルゴリズムの実装に関しては，Řehůřek ら[7]によって開発された Gensim を用いた．尚，平均や分散などの統計処理には numpy を用いている．

7.3. 解析対象

提案手法の解析対象データは Apple Inc. が提供している 2008 年 10 月 1 日から，2014 年 1 月 24 日までの Apple サポートコミュニティ[8]の質問文書，10391 文書を対象とした．

7.4. データクレンジング

対象が Web 上のデータのため解析するにあたり，データクレンジングを行った．まず記号等に関しては正規表現を用いて除去している．また Unicode 正規化と単語の小文字変換を用いて，文書に生成される単語の標準化を行った．加えてデータの可読性を向上させるため，1 文字の単語をストップワードとして除去している．

上記にて標準化を行った文書に対し，工藤らによって開発された形態素解析エンジン MeCab[9]を用いて形態素解析を行った．尚，各単語の活用形は標準形に変換している．また解析対象のデータがウェブ上のテキストデータであることを考慮し，MeCab の辞書に対しユーザ辞書として Wikipedia のタイトル語を追加している．

8. 実験結果

まず解析対象の文書に対して，LDA を用いた潜在トピックを抽出 KeyGraph による主張語の抽出を行った．

8.1. LDA による潜在トピックの抽出

LDA を用いて潜在トピックを抽出した（トピック数 T=30）．結果のうち，10 トピックに関して単語分布，人手で付与したラベルを表 1 に示す．

表 1　itune 関する潜在トピック

番号	単語分布	ラベル
TOPIC 1	ミュージック 表示 アップデート 設定 ios	ミュージック
TOPIC 2	アップデート バックアップ 復元 同期 アプリ	アップデート
TOPIC 3	認識 パソコン 復元 写真 表示	復元
TOPIC 4	再生 アプリ mac プレイリスト インストール	プレイリスト
TOPIC 5	連絡 アップデート 接続 データ pc	連絡データ 1
TOPIC 6	起動 バックアップ 写真 パソコン 購入	写真バックアップ
TOPIC 7	アップデート app 表示 接続 データ	アプリのアップデート
TOPIC 8	表示 エラー パソコン pc 画面	PC エラー
TOPIC 9	アプリ 接続 表示 音楽 エラー	音楽
TOPIC 10	アプリ 復元 データ pc 連絡	連絡データ 2

表 1 から確認できるように，LDA による潜在トピックでは，解析対象データから複数の話題が抽出されていることがわかる．

8.2. 土台の作成

次に提案手法で生成されたトピックを元に土台を作成する．土台を表 2 に示す．尚，重複した単語は生成確率の高いものを優先させた．

表 2 LDA を基とした土台

番号	単語
土台 1	ミュージック 表示 アップデート 設定 ios
土台 2	復元 同期 アプリ
土台 3	認識 パソコン 復元 写真
土台 4	再生 mac プレイリスト インストール
土台 5	データ pc
土台 6	起動 バックアップ 購入
土台 7	app 接続
土台 8	エラー 画面
土台 9	音楽
土台 10	連絡

8.3. KeyGraph の抽出

土台から得られた主張語を表 3 に示す.

表 3　主張語（Key 値上位 12 個まで）

番号	主張語	Key 値
主張語 1	itunes	0.98843
主張語 2	iphone	0.98814
主張語 3	お願い	0.963311
主張語 4	バックアップ	0.959242
主張語 5	ios	0.953329
主張語 6	同期	0.953306
主張語 7	復元	0.944833
主張語 8	状態	0.941808
主張語 9	表示	0.940347
主張語 10	使用	0.93915
主張語 11	方法	0.937182
主張語 12	データ	0.928066

主張 1～12 を確認したところ，トピックにも出現していたが，同期や復元・方法など，話題に関する内容が記載された主張が現れている．上位 12 個の Key では itunes,iphone など高頻度語が抽出多く抽出された.

表 4　主張語（Key 値上位 24 個まで）

番号	主張語	Key 値
主張語 13	アップデート	0.925759
主張語 14	アプリ	0.924016
主張語 15	問題	0.920832
主張語 16	ところ	0.919621
主張語 17	pc	0.910614
主張語 18	設定	0.908036
主張語 19	接続	0.906962
主張語 20	解決	0.902459
主張語 21	画面	0.901771
主張語 22	起動	0.900409
主張語 23	os	0.851203
主張語 24	どなた	0.844847

主張語 12～24 を確認したところ，接続・os など端末以外の箇所に関する内容や，問題・解決など質問への早期回答を求めるキーワードが抽出されていることがわかる．上位 24 個の Key では，土台に抽出されない主張語が抽出された.

尚，実験結果から，上位 24 個においても総じて Key 値が高い結果として抽出された.

8.4. 主張点ごとの平均評価極性値の比較

最後に iTune の話題を持ち，異なる主張点をもつ文書群同士の平均評価極性値の傾向に関して図 2 及び図 3 に示す.

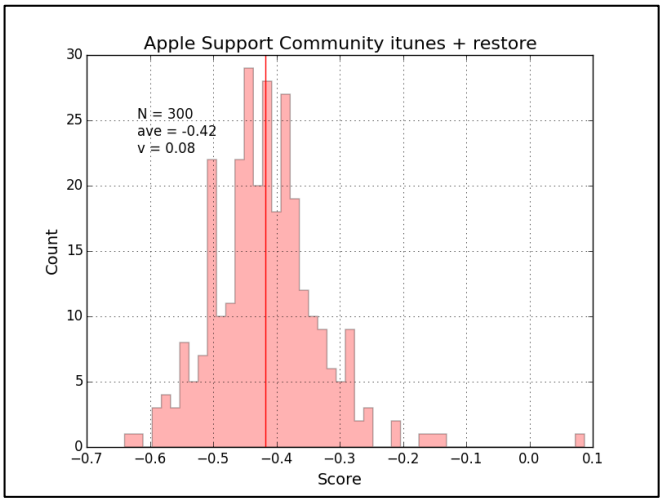


図 2　“itunes”＋“復元”

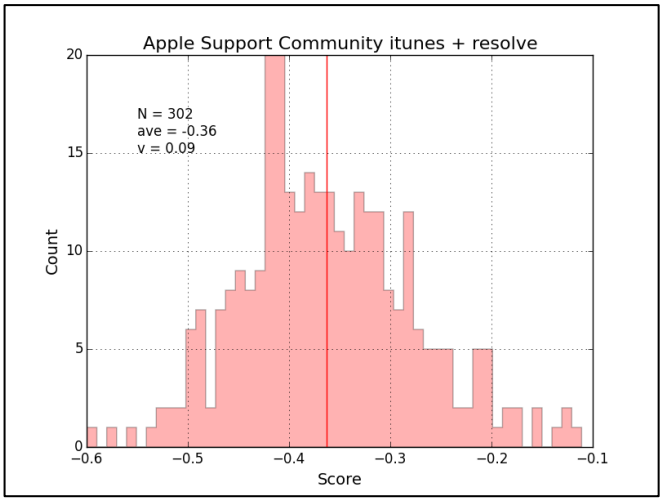


図 3　“itunes”＋“解決”

結果，異なる主張点を持つ文書群同士の比較では，平均評価極性値の差として，0.06 ほどの差異が確認できた．この結果，”復元”の主張を持つ文書群の方がよりネガティブな傾向を持つ文書群と判別された.

9. まとめ

今回は KeyGraph に LDA のトピックを適用させ，文書群の話題の主張語をとれるかを確認した.その結果，土台の主張語を抽出し，トピックに対する主張と見受けられる語が抽出し，同一トピック内で異なる主張を持つ文書群同士で評価極性に差を確認した.

今後は，文書に対して，トピックの混合比を基に主張を抽出し，それぞれのトピック別での主張語ごとを抽出したい．またそれを基に，文書の重要度をスコアリングし，人手評価と比較し，提案手法の有用性を評価したい.

参　考　文　献

[1] 輪島幸治，小河誠巳，古川利博，嶋田 茂，“潜在的ディリクレ配分法を用いたネガティブ要因分析，電子情報通信学会 他共催，第 6 回データ工学と情報マネジメントに関するフォーラム DEIM2014 論文集，A9-3 (Mar.　2014).

[1] Turney, Peter D., “Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews“, ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424,2002.

[2] 乾 孝司，奥村 学，"テキストを対象とした評価情 報の分析に関する研究動向", Journal of natural language processing 13(3)，201-241，2006-07-10.

[3] 輪島幸治，小河誠巳，古川利博，嶋田茂，” 共起語の評価極性に着目したネガティブトピックの評価”，電子情報通信学会技術研究報告 信学技報 114(101)，73-78，2014-06-21 電子情報通信学会.

[4] 大澤 幸生，ベルソン ネルス E，谷内田正彦，” 語の共起グラフの分割・統合によるキーワード抽出”，電子情報通信学会論文誌. D-I，情報・システム，I-コンピュータ J82-D-I(2)，391-400，1999-02-25

[5] Blei et al，“Latent Dirichlet Allocation”，Journal of Machine Learning Research，Vol. 3，pp. 993-1022，2003.

[6] Radim Řehůřek and Petr Sojka，Software Framework for Topic Modelling with Large Corpora，Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks，pp. 45-50, (May 22. 2010).

[7] アップルジャパン合同会社,"Apple サポートコミュニティ", https://discussionsjapan. apple. com/，最終閲覧日：2014/1/5.

[8] Taku Kudo，Kaoru Yamamoto，Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis，Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)，pp. 230-237 (2004.)

[9] Gephi, http://oss. infoscience. co. jp/gephi/gephi. org/，最終閲覧日：2015/1/13.

付　録

A1. 単語感情極性対応表の極性分布

単語感情極性対応表の極性傾向を評価したデータに関しては,ネガティブな極性結果が現れやすいため，単語感情極性対応表（日本語),（英語）の評価を実施した．評価対象には，本研究における解析対象データに加え，中立な文書群として日本語 Wikipedia の Abstract データのうち，適切に平均評価極性値が算出可能であった 677400 文書を対象としている．

単語感情極性対応表（日本語）（英語）の極性分布の算出結果に関して，図 4 および図 5 に示す．図は縦軸が単語の数，横軸はその単語が持つ極性値の値である．図 6 の結果から明らかなように，単語感情極性対応表（日本語）の分布に関しては，負値を持つ単語の数が非常に多く，負値に寄っている（μ =-0.32）ことが分かった．対して図 7 に関しては，正値・負値をもつ単語の数が平均的であるものの，単語の分布の範囲が非常に狭い分布の範囲が非常に狭い（σ =0.24）ことが分かった．

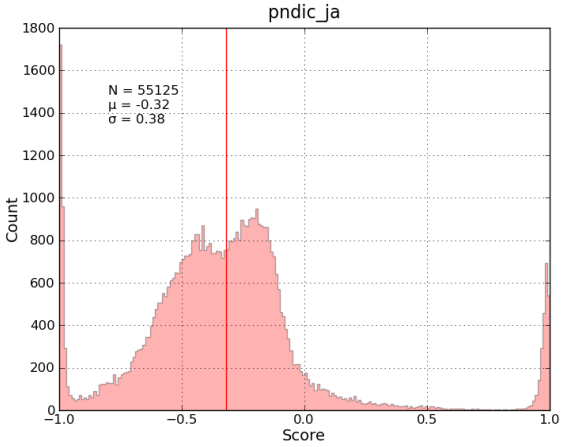


図 8　単語感情極性対応表（日本語）

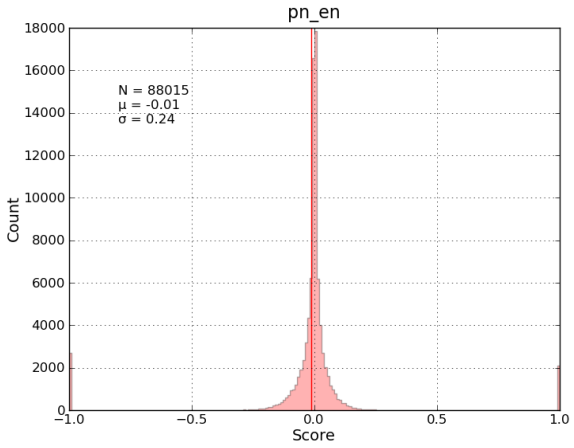


図 9　単語感情極性対応表（英語）

A2. 文書群の極性分布

単語感情極性対応表（日本語）を用いて，文書群を極性評価した歳の算出結果に関して，図 10 および図 11 に示す.

図は縦軸が文書数，横軸は文書を単語感情極性対応表（日本語）を用いて平均評価極性値を算出した平均極性値の値である.

評価の結果，解析対象データ，比較対象データともにヒストグラムの平均値が負値(各 μ =-0.43, μ =-0.35)となった．この結果から，単語感情極性対応表（日本語）を用いた場合，辞書の極性分布に依存し，対象データの平均評価極性値のほとんどは負値となってしまうことが明らかになった.

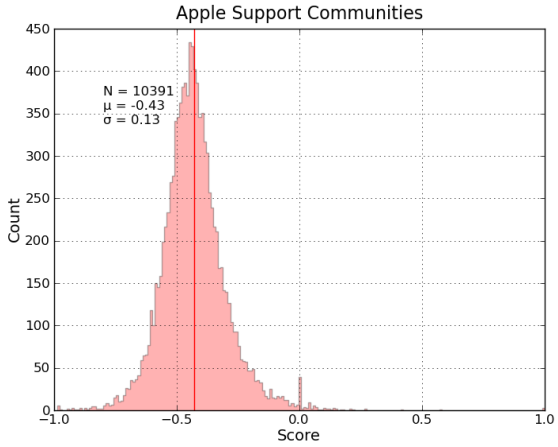


図 12　Apple サポートコミュニティ文書の評価結果

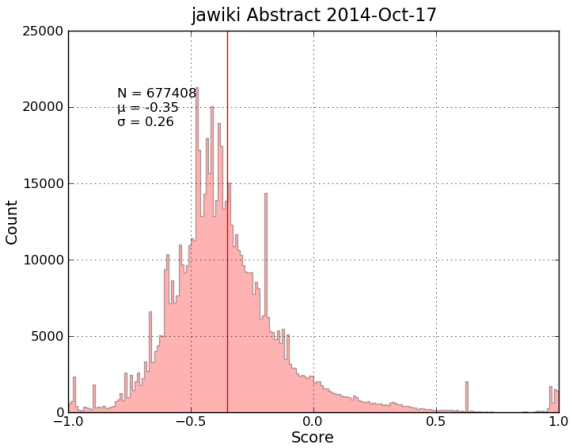


図 13　Wikipedia Abstract 文書の評価結果