Vol.2010-DBS-151 No.32
2010/11/13

Wikipediaにおけるエントリ粒度の評価

島田 諭[†1] 佐藤哲司[†1]

事典の品質や特性を決定づける指標として，エントリの粒度がある．ある概念が集約的に記述されている度合いと，概念間が相互に参照される度合いのバランスを意味し，事典の全体的な特性を示唆するとともに，個々のエントリの品質を反映すると考えられる．本論文では，ソフトウェアの評価指標である凝集度を応用し，事典型テキストにおけるエントリ粒度を測定する手法を提案する．事典のエントリとソフトウェアのモジュールを対応させ，出現語間の関係のモデル化を行う．Wikipediaのアブストラクトにおける出現語間の関係を，語の出現頻度の高低に着目して有向化し，エントリの凝集度を算出する．評価実験の結果，提案手法により安定したエントリ粒度の推定が行えること，算出される凝集度の値は，アブストラクトの文字列長では判別不可能な特性を反映していることが示唆され，提案手法の有用性が確認できた．

Evaluating Cohesion of Entries in the Wikipedia

Satoshi Shimada[†1] and Tetsuji Satoh[†1]

As an important index that can decide quality and characteristic of cyclopedia, there is cohesion of entries. It is thought that it means a balance of intensiveness and relativeness of each entries, and that it reflects the overall characteristic of cyclopedia and the quality of an individual entry. In this paper, we propose a estimating method of cohesion of entries in cyclopedic texts applying cohesion metrics used to evaluate maintainability of software modules. We have an entry of cyclopedia correspond to a module of software, and we define the model of relations between appearance words in entries. Those relations are directed based on order of frequency, and the cohesion of the entry is calculated. As a result, we confirmed that the cohesion is stably calculated using our method, and that calculated values reflect latent characteristics of entries against character length.

†1 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba

1. はじめに

近年，Wikipediaは盛んに利用されているが，その品質については議論が続いている．事典の検索性や読みやすさに影響する指標として，エントリ（項目）数がある．Wikipediaのエントリ数は，市販の百科事典の数倍から10倍程度であり，内容が多岐にわたっている．Wikipedia日本語版では，2010年8月14日現在，約690,000のエントリが存在する[⋆1]．例えば，平凡社の「改訂新版世界大百科事典」では約420,000（大項目数は約90,000）[⋆2],「ブリタニカ国際百科事典小項目版2010」では約138,000である[⋆3]のと比較し，Wikipediaのエントリ数は群を抜いて多く，また，現時点でも増加し続けている．学習用のマルチメディア百科事典である「Microsoft エンカルタ総合大百科2007」のエントリ数は約36,000であり，20倍以上の差がある[⋆4]．このように，取り上げる分野や対象とする読者によって必要となるエントリ数は異なるため，単純にエントリ数の多寡によって事典の品質を推定することはできない．

一方，辞書学（lexicography）では，エントリの粒度が，事典の品質や特性を決定づける重要な指標であると考えられている．エントリ粒度は，ある概念が集約的に記述されている度合いと，概念間が相互に参照される度合いのバランスを意味する．エントリ粒度の分布が事典の全体的な特性を示唆するだけでなく，個々のエントリの粒度は，そのエントリの品質を反映すると考えられる．

本論文では，ソフトウェアの評価指標である凝集度を応用し，事典型テキストにおけるエントリ粒度を測定する手法を提案する．凝集度は，モジュール間でのメソッドとメンバ変数の機能的協調度合いを測定する指標である．事典のエントリとソフトウェアのモジュールを対応させ，出現語間の関係を用いることで，凝集度を事典型テキストに適用できると考えられる．提案手法では，Wikipediaのアブストラクトにおける出現語間の関係を，語の出現頻度の高低に基づき有向化し，凝集度を算出する．算出される値と，エントリの文字列長および異なり出現語数との相関を調べ，エントリ粒度の評価指標としての妥当性を評価する．

以下，2章で関連研究を概説し本研究の位置付けを述べ，3章で提案手法を詳説し，4章で評価実験の結果について述べ，5章で考察し，6章でまとめと今後の展望について述べる．

⋆1 http://ja.wikipedia.org/
⋆2 http://www.heibonsha.co.jp/catalogue/exec/frame.cgi?page=browse.cgi&code=034990
⋆3 http://www.britannica.co.jp/products/encyclopedia/logovista2010.html
⋆4 http://www.microsoft.com/japan/presspass/detail.aspx?newsid=2820

2. 関 連 研 究

Wikipedia における情報の品質（IQ：Information Quality）は，様々な側面から検討されている．Stvilia らは，編集回数などのメタデータに基づいて，エントリの質を 7 種類の側面から推定する手法を提案している[1)]．Emigh らは，分野別にエントリの語数と文字列長を分析している[2)]．Dondio らは，編集回数の多い上位投稿者に着目し，エントリの内容の安定度から内容の信頼性を推定する手法を提案している[3)]．Hu らは，エントリの出現語とその投稿者を対応させた二部グラフを構築し，投稿者のオーソリティ値からエントリの品質を推定する手法を提案している[4)]．

Wikipedia のリンク構造を用いる従来研究では，エントリ間に有用なリンク構造があることを前提としている点で共通している．しかし，Wikipedia におけるエントリ間のリンクは，基本的に人手で付与されている．不特定多数のユーザが編集作業を担う Wikipedia において，事典全体でのバランスや使いやすさを考慮してリンク構造を構築することは，きわめて困難である．実際に，関連するエントリ間で一方向にしかリンクがない，代表的とはいえない些末なエントリへのリンクがあるといった状況がある．現状では，事典の規模に見合う十分な量と質を備えたリンク構造が構築されているとは言いがたく，品質推定の根拠とするには不十分である．

一方，文書集合の共起語グラフは一般に，人手で付与されたリンク構造よりも密となる．共起語グラフを用いることにより，エントリ間の関連性と，個々のエントリの内容の両方を考慮してエントリ粒度を評価できると考えられる．なお，Wikipedia ではエントリの文字数や出現語数の分散が大きいと予想されるため，本研究ではアブストラクトを用いる．アブストラクトとは，エントリの冒頭部分（リード）のテキストをいう．中山らは，利用者の目につきやすい Wikipedia のアブストラクトは，よく編集される傾向があることを指摘し，特にシソーラスの構築に有用な意味関係がアブストラクトに多く含まれることを報告している[5)]．アブストラクトは，エントリ全体の特性を反映しながらも，エントリの文字数にかかわらず文字数の分散が小さいことが予想されることから，エントリの文字数に影響されずにエントリ粒度を評価するために有用であると考える．

3. 出現語間の関係に基づくエントリ粒度の推定手法

3.1 凝集度の事典型テキストへの応用

本研究では，これまでソフトウェアの品質管理に用いられてきた凝集度を応用し，Wikipedia に代表される Web 上の事典型テキストにおけるエントリの粒度を推定する手法を提案する．凝集度とは，ソフトウェアモジュール内における情報要素（メンバ変数）と機能要素（メソッド）の間の関連度を示す指標である．凝集度が高い，すなわち無駄な変数やメソッドがなくそれぞれが分担する範囲が明確で，変数とメソッドが密接に関連しているモジュールは，コードの可読性に優れ，再利用性や保守性が高いとされている．

凝集度の指標として，Chidamber らの $LCOM$（Lack of Cohesion in Methods）[6)]，Bieman らの TCC（Tight Class Cohesion），LCC（Loose Class Cohesion）[7)] などが知られている．$LCOM$ を改良した Henderson-Sellers の $LCOM*$ は (1) 式で定義されている[8)]．

$$LCOM* = \frac{\frac{1}{a}\sum_j^a \mu(Aj) - m}{1 - m} \tag{1}$$

ここで，A_j は着目しているクラスの j 番目のメンバ変数，a はメンバ変数の数，m はメソッドの数，$\mu(A_j)$ はメンバ変数 A_j にアクセスしているメソッドの数である．

TCC は (2) 式，LCC は (3) 式で定義されている．

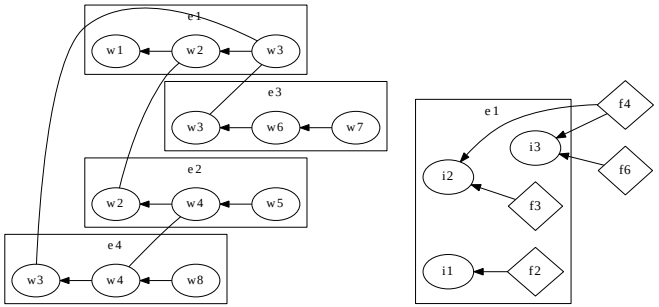
$$TCC(C) = \frac{NDC(C)}{NP(C)} \tag{2}$$

$$LCC(C) = \frac{NDC(C) + NIC(C)}{NP(C)} \tag{3}$$

ここで，着目するクラス C のメソッドの個数を N とする．$NP(C)$ はクラス中でのメソッド間の最大接続可能数であり，$N*(N-1)/2$ となる．NDC は，メソッド間の直接接続数（Number of Direct Connections），NIC は間接接続数（Number of Indirect Connections）である．すなわち，TCC は接続の密度（connection density），LCC は全体的な接続性（overall connectedness）を意味する．

$LCOM*$ は，評価する要素と直接関係する範囲のみに着目し，局所的な凝集度を算出するのに対し，TCC および LCC は，間接的に接続された要素にも着目し，大域的な凝集度を算出する．要素間の依存関係の最小化を前提とするソフトウェアとは異なり，事典型テキストにおいてはエントリ間に密なグラフが生成されることを前提とする．共起語グラフは一般に，平均距離が短く平均クラスター係数が非常に高い Small-world グラフとなることが知られている[9)]．大部分のノードが間接的に接続された状態になり LCC の差が小さくなると予想されることから，エントリ粒度を推定する尺度として LCC は適さないと考えられる．このため，本研究では，局所的な凝集度を算出する $LCOM*$ を拡張し，事典型テキストのエントリ粒度の推定への適用性を検討する．

Vol.2010-DBS-151 No.32
2010/11/13



(a) エントリ内およびエントリ間における関係

(b) 情報要素と機能要素に分類した関係

図 1 エントリにおける出現語間の関係のモデル
Fig. 1 The model of word's relation in proposal method.

凝集度を用いてオントロジの品質を推定する手法は以前から提案されている[10] が，近年は，品質のばらつきが大きい Web 上のテキストを用いたオントロジに対し凝集度を適用する手法が注目されている[11]．本研究は，オントロジではなく自然文である事典型テキストを対象とし，エントリのアブストラクトにおける出現語間の関係を用いてエントリの凝集度を算出する点で，先行研究とは異なる．

3.2 モ デ ル

凝集度を応用し，出現語間の関係を用いてエントリ粒度を算出するためには，(1) 凝集度における情報要素（メンバ変数）と機能要素（メソッド）に相当する要素を定義すること，(2) 要素間の関係を有向化できることが要件となる．本論文では，(1) 文書中の出現語が，情報要素と機能要素を兼ねるとみなし，(2) 出現語間の関係を，語の出現頻度の高低に基づき有向化する．

文書中の出現語には，情報要素としての側面と，機能要素としての側面がある．情報要素としての側面とは，ある語を用いることによって，その語が表現する内容を示すことである．機能要素としての側面とは，他の文書と共通の語を用いることによって，共通の語が出現する文書間の関連を示すことである．

エントリにおける出現語間の関係のモデルを**図 1** に示す．**図 1(a)** において，e_n はエントリ，w_k は出現語を表わす．出現語間の関係には，同一のエントリ内で共起する関係（例えば，図中 e_1 における w_1，w_2，w_3），および複数のエントリ間で共通の語が用いられるという関係（同，図中 e_1 と e_2 における w_2）の 2 種類がある．さらに，e_3 において w_3 を用いることが，e_1 において共起する w_2，w_1 を読者に連想させる働きがあるとすれば，ソフトウェアにおいて機能要素が情報要素にアクセスする関係に相当すると考えられる．また，このような働きが，語の出現頻度の高低によって制約を受けると仮定すれば，出現語の共起関係を有向化できる．ここで，k を語の出現頻度の順位とし，エントリ内で共起関係にある 2 語間で，順位が低い（k が大きい）語を機能要素，高い（k が小さい）語を情報要素とみなすことができる．**図 1(b)** において，f_m は機能要素，i_j は情報要素である．エントリ e_1 の凝集度を算出するには，f_2，f_3，f_4，f_6 と，i_1，i_2，i_3 の関係を用いる．

3.3 出現語間の関係の有向化手法

表面的には無向である出現語間の関係を有向化するには，文中での語の出現位置，出現頻度や重要度の高低などが利用できると考えられる．本論文では，基礎的な統計量である文書頻度 *df* を用い，各エントリのアブストラクトにおいて，着目する語とその他の語の *df* を比較し，以下の 4 通りの方法で有向の共起関係があるとみなす．

(1) 着目する語から，その語よりも *df* が低いすべて語に対し，有向の共起関係があるとみなす．着目する語の，すべてのエントリにおける異なり共起語数を deg_L とする．
(2) 着目する語から，その語よりも *df* が 1 段階低い語に対し，有向の共起関係があるとみなす．着目する語の，すべてのエントリにおける異なり共起語数を deg_{L1} とする．
(3) 着目する語から，その語よりも *df* が 1 段階高い語に対し，有向の共起関係があるとみなす．着目する語の，すべてのエントリにおける異なり共起語数を deg_{H1} とする．
(4) 着目する語から，その語よりも *df* が高いすべて語に対し，有向の共起関係があるとみなす．着目する語の，すべてのエントリにおける異なり共起語数を deg_H とする．

4. 評　　価

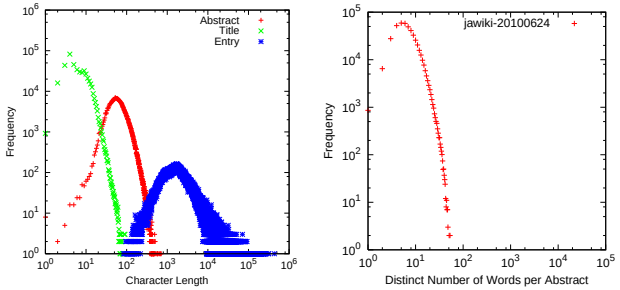
4.1 実 験 概 要

本実験では，まず，共起語数 deg_L，deg_{L1}，deg_{H1}，deg_H の，文書集合内での分布を明らかにし，これらの値および *df* との相関を明らかにする．次に，deg_L，deg_{L1}，deg_{H1}，deg_H の各値を用いた凝集度，すなわち，$LCOM{*_L}$，$LCOM{*_{L1}}$，$LCOM{*_{H1}}$，$LCOM{*_H}$ を算出し，これらの値の分布およびエントリの文字列長および異なり出現語数との相関を明らかにする．

Vol.2010-DBS-151 No.32
2010/11/13

表 1 Wikipedia のダンプデータから抽出されたエントリ数
Table 1 Number of entries extracted from dumped data of the Wikipedia.

条件	件数	
アブストラクトの末尾が句点「。」	**447,882**	(65.7%)
アブストラクトが空白	102,419	(15.0%)
(上記以外)	131,862	(19.3%)
(全体)	682,163	



(a) 文字列長　(b) エントリにおける異なり出現語数

図 2 エントリ，タイトル，アブストラクトにおける文字列長および異なり出現語数の分布
Fig. 2 Distributions of character length in entries, title and abstract, and distinct number of words.

表 2 エントリ，タイトル，アブストラクトにおける文字列長の平均，分散，中央値，最小値，最大値
Table 2 Basic statistics of character length in entries, titles and abstracts.

	平均	分散	中央値	最小値	最大値
エントリ	4,961.9	7.3×10^7	2,654	85	433,079
タイトル	7.9	2.7×10^1	7	1	91
アブストラクト	76.9	1.7×10^3	67	1	707

実験に用いる事典型テキスト

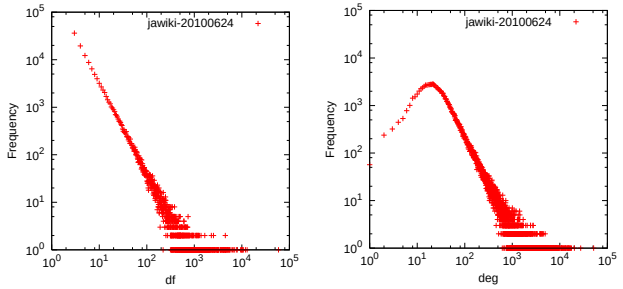
本実験では，Wikipedia 日本語版のアブストラクトを用いて評価実験を行う[⋆1]．XML 構文，および Wikipedia の構築に使用されている MediaWiki の構文を処理するパーサを Perl で実装し，このパーサを用いて，エントリのタイトルおよびアブストラクトのテキストを抽出した．抽出されたエントリ数を**表 1** に示す．抽出されたエントリのうち，アブストラクトの末尾が句点「。」でないエントリ，タイトルが「曖昧さ回避」を含むまたは「一覧」で終わるエントリ，アブストラクトに「一覧」を含むエントリ，アブストラクトが空白になっているエントリを除外し，447,882 のエントリを実験に用いる．

これらのアブストラクトから，2 文字以上の漢字またはカタカナからなる文字列，および，3 文字以上の英数字からなる文字列を語として抽出した．アブストラクトにおけるエントリあたり出現語数の分布を**図 2(b)** に示す．$df \geq 3$ となる出現語の異なり数は 128,521 である．なお，$df < 3$ となる出現語の異なり数は約 764,000 である．

⋆1 http://download.wikimedia.org/jawiki/20100624/jawiki-20100624-abstract.xml

表 3 エントリ，タイトル，アブストラクトにおける文字列長の相関係数
Table 3 A correlation coefficient matrix of character length in entries, titles and abstracts.

	エントリ	タイトル	アブストラクト
エントリ	1.000	0.060	**0.202**
タイトル		1.000	**0.246**
アブストラクト			1.000



(a) df　(b) deg

図 3 アブストラクトにおける出現語の文書頻度 df および共起語数 deg の分布
Fig. 3 Distributions of df and deg of words appeared in abstracts.

Vol.2010-DBS-151 No.32
2010/11/13

表 4 アブストラクトにおける出現語の文書頻度 df，共起語数 deg，deg_L，deg_{L1}，deg_{H1}，deg_H の相関係数
Table 4 A correlation coefficient matrix between df, deg, deg_L, deg_{L1}, deg_{H1}, deg_H of words appeared in abstracts.

	df	deg	deg_L	deg_{L1}	deg_{H1}	deg_H
df	1.000	0.796	0.818	0.688	**0.330**	**0.127**
deg		1.000	0.992	0.927	0.605	0.416
deg_L			1.000	0.901	0.509	**0.294**
deg_{L1}				1.000	0.743	0.509
deg_{H1}					1.000	0.854
deg_H						1.000

エントリ，タイトル，アブストラクトにおける文字列長の分布を**図 2(a)** に示す[⋆1⋆2]．いずれも対数正規分布を示し，タイトル，アブストラクト，エントリの順に文字列長が長くなっている．また，これらの平均，分散，中央値，最小値，最大値を**表 2** に示す．エントリの文字列長の分散は，アブストラクトの文字列長の分散より 4 桁大きい．また，これらの相関係数を**表 3** に示す．いずれの値の間でも，相関が非常に弱い．

4.2 実 験 結 果

共起語数 deg_L，deg_{L1}，deg_{H1}，deg_H の算出

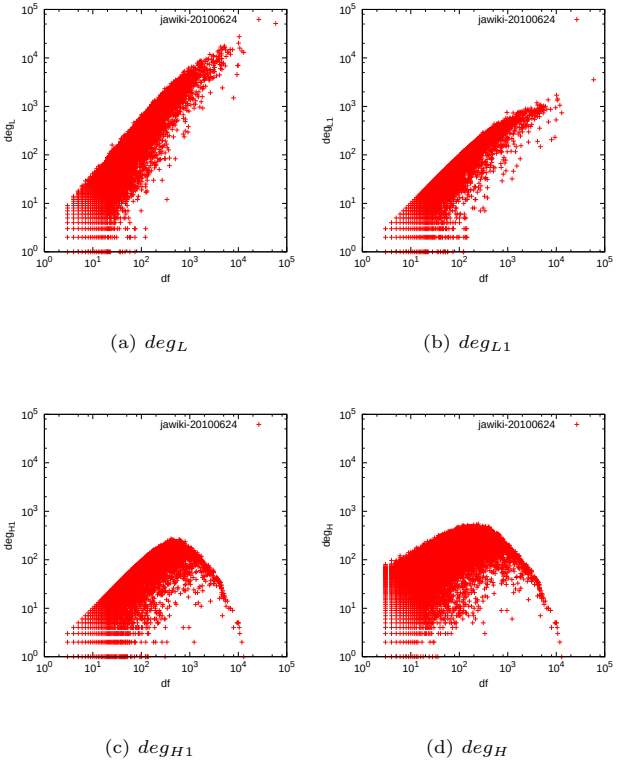
アブストラクトにおける出現語の文書頻度 df および共起語数 deg の分布を**図 3** に示す．df はベキ分布を示し，deg は対数正規分布を示す．deg の分布は，不特定多数のユーザが編集するため語彙が統制されにくいという Wikipedia の特性を反映している．

アブストラクトにおける出現語の共起語数 deg_L，deg_{L1}，deg_{H1}，deg_H を算出した．df に対する deg_L，deg_{L1}，deg_{H1}，deg_H の分布を**図 4** に示す．deg_L および deg_{L1} は，df に比例して単調増加している．deg_{H1}，deg_H は，対数正規分布を示している．また，これらの相関係数を**表 4** に示す．共起語数は，基本的には df との相関が強いが，deg_H および deg_{H1} だけは，その他の値との相関が弱い．特に，deg_H と df の相関係数は 0.127 であり，相関が非常に弱い．

共起語数 deg_L，deg_{L1}，deg_{H1}，deg_H を用いた凝集度 $LCOM*$ の算出

アブストラクトにおける出現語の deg_L，deg_{L1}，deg_{H1}，deg_H の値を用いて，(1) 式により，個々のエントリの凝集度 $LCOM*_L$，$LCOM*_{L1}$，$LCOM*_{H1}$，$LCOM*_H$ を算出

⋆1 マルチバイト文字は 1 文字と数えた．
⋆2 エントリの文字列長は，jawiki-20100624-page.sql.gz に含まれる page_len の値を用いた．

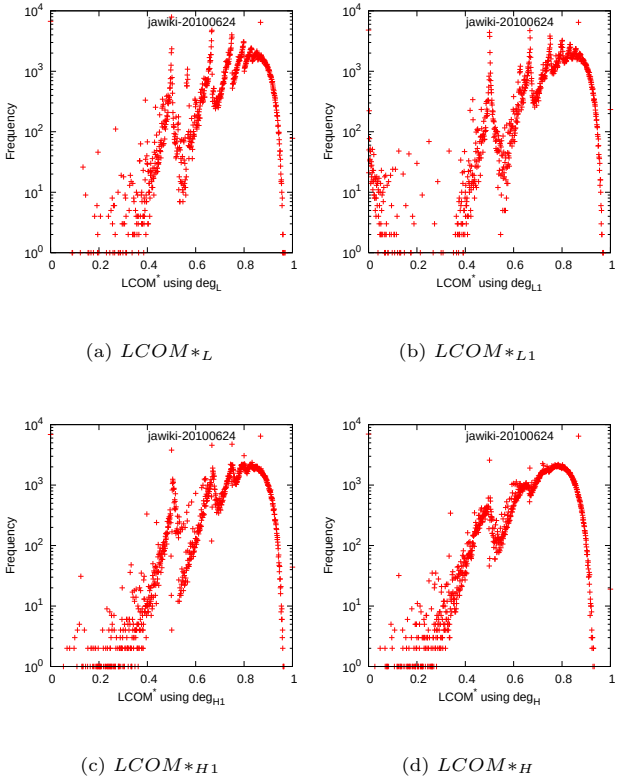


(a) deg_L (b) deg_{L1} (c) deg_{H1} (d) deg_H

図 4 アブストラクトにおける出現語の文書頻度 df に対する共起語数 deg_L，deg_{L1}，deg_{H1}，deg_H の分布
Fig. 4 Distributions of df and deg_L, deg_{L1}, deg_{H1}, deg_H of each words appered in abstreacts.

した．すなわち，着目するエントリ内の異なり出現語数を a，各々の出現語の共起語数を $\mu(A_j)$，着目するエントリ内のすべての出現語が持つ，全エントリ内での共起語の異なり数を m とし，$\mu(A_j)$ および m の算出に際し，deg_L，deg_{L1}，deg_{H1}，deg_H を使い分けることにより，4 種類の $LCOM*$ 値を得た．なお，異なり出現語数が 3 以上のエントリ 438,834 件を用いた．各 $LCOM*$ 値の分布を**図 5** に示す．いずれも，0.8 付近に頻度のピークがあり，0.97 付近で頻度が 0 となる．$LCOM*$ 値の低い領域に着目すると，$LCOM*_L$，$LCOM*_{L1}$，$LCOM*_{H1}$ は 0.4 付近で頻度が 0 に近づいているが，$LCOM*_H$ は 0.3 付近までなだらか

Vol.2010-DBS-151 No.32
2010/11/13



(a) $LCOM*_L$　(b) $LCOM*_{L1}$

(c) $LCOM*_{H1}$　(d) $LCOM*_H$

図 5　エントリの凝集度 $LCOM*_L$，$LCOM*_{L1}$，$LCOM*_{H1}$，$LCOM*_H$ の分布
Fig. 5　Distributions of $LCOM*_L$, $LCOM*_{L1}$, $LCOM*_{H1}$, $LCOM*_H$ of each entries.

に分布している．$LCOM*_{L1}$ に限り，0 から 0.1 の領域で頻度が高い．

エントリの文字列長および異なり出現語数と，各 $LCOM*$ 値の相関係数を**表 5** に示す．各 $LCOM*$ 値の間の相関係数は最小でも 0.957 であり，相関が非常に強い．一方，文字列長および異なり出現語数と，各 $LCOM*$ 値の間の相関は最大でも 0.619 であり，相関が弱い．特に，$LCOM*_H$ は文字列長との相関係数が 0.473 であり，相関係数が最小である．

$LCOM*_H$ の分布では，ピークは 0.784 であり，エントリ数は 2,188 である．頻度の落ち込みが見られる 0.499 でのエントリ数は 46 である．凝集度の 0.3 以下の領域で最も頻度の高い 0.278 でのエントリ数は 35 である．

表 5　アブストラクトの文字列長，異なり出現語数とエントリの凝集度 $LCOM*$ の相関係数
Table 5　A correlation coefficient matrix between character length of abstracts, distinct number of words in abstracts, and $LCOM*$ of entries.

	文字列長	異なり出現語数	$LCOM*_L$	$LCOM*_{L1}$	$LCOM*_{H1}$	$LCOM*_H$
文字列長	1.000	0.855	0.501	0.496	0.478	**0.473**
異なり出現語数		1.000	0.619	0.613	0.593	**0.591**
$LCOM*_L$			1.000	0.980	0.970	0.959
$LCOM*_{L1}$				1.000	0.977	0.957
$LCOM*_{H1}$					1.000	0.986
$LCOM*_H$						1.000

5. 考　　察

本論文では，ソフトウェアの評価指標である凝集度の，事典型テキストにおけるエントリ粒度の推定への適用性を検討するための評価実験を行った．以下では，実験項目に対応して考察する．

事典型テキストのエントリ粒度推定において，アブストラクトを用いることの妥当性を検討するため，エントリおよびアブストラクトにおける文字列長の分析を行った．その結果，エントリ全体の文字列長とアブストラクトの文字列長に相関はほとんどなく，これらとエントリのタイトルの文字列長との間にも相関はほとんどなかった．実際のエントリとアブストラクトを見ると，エントリの文字列長は分野や内容によって大きく異なり，必ずしも単純に文字列長によって読みやすさが左右されるのではないことがわかった．一方で，アブストラクトの文字列長が長いエントリでは，アブストラクトでの説明に固有表現が多用されている，エントリ内での説明項目の分け方が明確でないなどの傾向が見られた．また，文字列長の分散は，エントリでは大きいがアブストラクトでは小さいことから，アブストラクトを用いることで安定したエントリ粒度の推定が行えるという見通しが得られた．

凝集度の算出において出現語間の関係を有向化する手法を評価するため，出現語の共起語数 deg_L，deg_{L1}，deg_{H1}，deg_H の算出および分析を行った．deg_L，deg_{L1} は，df に比例して単調増加する分布を示し，deg_{H1}，deg_H は，df に対して対数正規分布を示すことがわかった．このことから，df が低い語から高い語へのリンクを生成する deg_H においては，df がきわめて高い一般語の影響が少なく，局所的な凝集度の算出に適すると考えられる．ま

た，df，deg との相関は，deg_L で最も強く，deg_H で最も弱くなることがわかった．以上の結果から，提案手法では，出現語間の関係の有向化に df を用いながらも，一般に文書集合中での変動が大きいとされる語の df そのものによる影響が少なく，安定してエントリ粒度の推定が行えることが期待される．

次に，deg_L，deg_{L1}，deg_{H1}，deg_H を用いて，凝集度 $LCOM*_L$，$LCOM*_{L1}$，$LCOM*_{H1}$，$LCOM*_H$ を算出し，分析を行った．各 $LCOM*$ 値は，ほぼ同様の分布を示し，これらの値の間の相関が非常に強いことがわかった．ただし，$LCOM*_{L1}$ では，値の低い領域での分布が，他の 3 種類の値とは異なっていた．一方，これらの値とアブストラクトの文字列長および異なり出現語数との相関係数は 0.5 から 0.6 程度の範囲となり，弱い相関があることがわかった．このことから，算出した各 $LCOM*$ 値は，エントリの特性のうち，アブストラクトの文字列長によって判別可能な特性と，文字列長では判別不可能な特性の両方を反映していると考えられる．

以下では，実際のエントリの内容に着目して考察する．

Wikipedia では，内容が煩雑なエントリに対する「分割提案」，および内容が少なく独立したエントリとすることが適当でないエントリに対する「統合提案」をユーザが行えるようになっている[⋆1]．このような提案がなされたエントリにおける文字列長と，そのアブストラクトにおける異なり出現語数，およびエントリの各 $LCOM*$ 値を抜粋して**表 6** に示す．表中に示した種別「分割提案」および「統合提案」は，2010 年 10 月 18 日時点でこれらの提案がなされているエントリである．「分割済み」は，2010 年 1 月から 6 月までの間にこれらの提案がなされ，実験に用いたデータの取得日時である 2010 年 6 月 24 日までに分割が行われたエントリである．

「分割提案」では，文字列長の最大と最小で 2 桁差がある．$LCOM*$ 値の最大は「仮面ライダー W」の $LCOM*_{L1}$ で 0.935，最小は「京都国体」の $LCOM*_H$ で 0.682 だった．実際のエントリを見ると，「仮面ライダー W」は内容が煩雑であり，アブストラクトの異なり出現語数も比較的多い．「指定区間」はエントリの文字列長は比較的短いが，アブストラクトの内容が煩雑だった．「インターネットオークション」は，エントリ内のごく一部を別のエントリに転記するよう提案されているだけであり，エントリ自体には特段の問題は見られなかった．「国際ターミナル駅」「京都国体」は，きわめて簡潔に書かれており，分割すると細分化しすぎるように見受けられた．これらのエントリの特性と $LCOM*$ の値を対応させると，分割の必要性が高いと思われるエントリの $LCOM*$ 値は 0.9 前後であるのに対し，分割の必要性が低いと思われるエントリの $LCOM*$ 値は 0.7 前後になるという傾向が認められる．また，「仮面ライダー W」と「指定区間」の文字列長が大きく異なることから，文字列長では判別できないエントリの特性を $LCOM*$ 値が反映していると考えられる．

「分割済み」では，浪曲の名跡である「相模太郎」と，特定の人物の説明である「相模太郎 (初代)」において，$LCOM*$ 値が比較的低くなっている．また，「仮面ライダー W」と同程度に文字列長が長い「ポケットモンスター SPECIAL の登場人物」においても，$LCOM*$ 値は「仮面ライダー W」よりも低い 0.8 程度となっている．これらの結果から，エントリが繰り返し編集されて洗練されたり，適切な分割が行われることにより，$LCOM*$ 値が低下，すなわち凝集度が高まると考えられる．

一方，「統合提案」では，$LCOM*$ 値が 0.5 前後から 0.9 前後まで幅広く分布していた．「ラフォーレ原宿アートワーク」と「ワンツーパンチ」は，エントリの文字列長がきわめて短いだけでなく，単独でエントリとすることが妥当とは言いがたい内容になっていた．これらの $LCOM*$ 値は比較的低く，特に「ラフォーレ原宿アートワーク」の $LCOM*_H$ は 0.423 で

表 6 分割または統合が提案されたエントリにおける文字列長，異なり出現語数および凝集度 $LCOM*$
Table 6 A list of character length of entries, distinct number of words in abstracts and $LCOM*$ of entries proposed to divide or merge on the Wikipedia.

種別		異なり	$LCOM*$			
タイトル	文字列長	出現語数	L	$L1$	$H1$	H
分割提案						
仮面ライダー W	165,289	21	0.933	0.935	0.927	0.882
インターネットオークション	20,927	8	0.798	0.794	0.808	0.783
国際ターミナル駅	9,603	6	0.746	0.764	0.753	0.704
指定区間	4,480	16	0.902	0.903	0.881	0.842
京都国体	2,548	7	0.721	0.725	0.716	0.682
分割済み						
ポケットモンスター SPECIAL の登場人物	142,821	7	0.794	0.811	0.793	0.752
相模太郎 (初代)	1,157	3	0.500	0.500	0.522	0.510
相模太郎	220	4	0.742	0.755	0.746	0.712
統合提案						
元禄赤穂事件	98,177	7	0.830	0.832	0.822	0.775
機種依存文字	12,171	21	0.918	0.920	0.908	0.859
ふるさと大使	5,726	12	0.908	0.914	0.905	0.855
観光大使	2,402	12	0.897	0.904	0.890	0.825
ラフォーレ原宿アートワーク	696	5	0.498	0.503	0.502	0.423
ワンツーパンチ	298	5	0.670	0.699	0.681	0.649

⋆1 http://ja.wikipedia.org/wiki/Category:依頼と提案

あり，表 6 中で最小である．このことから，要素間の依存関係の最小化を前提とするソフトウェアとは異なり，エントリ間に密な関係が構築されることを前提とする事典型テキストにおいては，$LCOM*$ の値が小さいほど優れたエントリであるとは限らないことがわかった．むしろ，凝集度が低下しすぎない範囲で，ある程度高い $LCOM*$ 値を示すエントリを，最適な粒度のエントリとみなすといった，ソフトウェアにおける凝集度とは異なる値の解釈をする必要があることが示唆された．

6. おわりに

本論文では，ソフトウェアの評価指標である凝集度を応用し，事典型テキストにおけるエントリの粒度を測定する手法を提案した．凝集度は，モジュール間でのメソッドとメンバ変数の機能的協調度合いを測定する指標である．事典のエントリとソフトウェアのモジュールを対応させ，出現語間の関係を用いることで，凝集度を事典型テキストに適用できると考え，出現語間の関係のモデル化を行った．

提案手法では，Wikipedia のアブストラクトにおける出現語間の関係を，語の出現頻度の高低に基づき有向化し，凝集度を算出する．事典型テキストのエントリ粒度推定において，アブストラクトを用いることの妥当性を検討するため，エントリおよびアブストラクトにおける文字列長の分析を行った結果，アブストラクトを用いることで安定したエントリ粒度の推定が行えることが示唆された．

凝集度の算出において出現語間の関係を有向化する手法を評価するため，出現語の共起語数 deg_L，deg_{L1}，deg_{H1}，deg_H の算出および分析を行った．その結果，提案手法では，出現語間の関係の有向化に df を用いながらも，一般に文書集合中での変動が大きいとされる語の df そのものによる影響が少なく，安定してエントリ粒度の推定が行えるという見通しが得られた．特に，df が低い語から高い語へのリンクを生成する deg_H においては，df がきわめて高い一般語の影響が少なく，局所的な凝集度の算出に適すると期待できる．

次に，共起語数 deg_L，deg_{L1}，deg_{H1}，deg_H を用いて，凝集度 $LCOM*_L$，$LCOM*_{L1}$，$LCOM*_{H1}$，$LCOM*_H$ を算出し，分析を行った．その結果，算出した凝集度の値は，事典型テキストのエントリが示す特性のうち，アブストラクトの文字列長によって判別可能な特性と，文字列長では判別不可能な特性の両方を反映していると考えられることがわかり，提案手法の有用性が確認できた．

今後の課題として，出現語間の関係の有向化に出現頻度以外の特徴量を用いた場合の評価や，エントリの文字列長と $LCOM*$ 値の両方を考慮したエントリ粒度の評価，より大域的なエントリ粒度の評価への対応などが考えられる．提案手法の応用としては，Wikipedia における「分割提案」および「統合提案」の自動化，編集により変化するエントリ粒度の予測などが考えられる．

謝辞 本研究は科研費（21500091）の助成を受けたものである．ここに記し謝意を示す．

参 考 文 献

1) Stvilia, B., Twidale, M. B., Smith, L. C. and Gasser, L.: Assessing information quality of a community-based encyclopedia, *In Proceedings of the International Conference on Information Quality*, pp.442–454 (2005).
2) Emigh, W. and Herring, S.: Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias, *System Sciences, 2005. HICSS '05. Proceedings of the 38th Annual Hawaii International Conference on*, pp.99a – 99a (2005).
3) Dondio, P. and Barrett, S.: Computational Trust in Web Content Quality: A Comparative Evaluation on the Wikipedia Project, *Informatica*, pp.151–160 (2007).
4) Hu, M., Lim, E.-P., Sun, A., Lauw, H. W. and Vuong, B.-Q.: Measuring article quality in wikipedia: models and evaluation, *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, New York, NY, USA, ACM, pp.243–252 (2007).
5) 中山浩太郎，原　隆浩，西尾章治郎：自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築，日本データベース学会論文誌，Vol. 7, No.1, pp.67–72 (2008).
6) Chidamber, S. and Kamerer, C.: A metrics Suite for Object-Oriented Design, Vol.SE-20, No.6 (1994).
7) Bieman, J.M. and Kang, B.-K.: Cohesion and reuse in an object-oriented system, *SSR '95: Proceedings of the 1995 Symposium on Software reusability*, New York, NY, USA, ACM, pp.259–262 (1995).
8) Henderson-Sellers, B.: *Object-oriented metrics: measures of complexity*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1996).
9) Ferrer, R. and Sole, R. V.: The small world of human language, *Proceedings of The Royal Society of London. Series B, Biological Sciences*, Vol.268, pp.2261–2265 (2001).
10) 堀　雅洋：結合度と凝集度に基づくオントロジーの評価，知識ベースシステム研究会，pp.39–44 (1998).
11) Ma, Y., Jin, B. and Feng, Y.: Semantic oriented ontology cohesion metrics for ontology-based systems, *Journal of Systems and Software*, Vol. 83, No. 1, pp. 143–152 (2010).