# Wikipediaにおけるミッシングリンクの自動発見手法

中川 隆人[†1]　古瀬 一隆[†2]　陳 漢雄[†3]

[†1, †2, †3, †4] 筑波大学大学院システム情報工学研究科

## 1．はじめに

近年のインターネットの普及により WWW 上のコンテンツ量は飛躍的に増大してきた。近年特定の情報を得るために Wikipedia[1][2]を使うユーザも多くなってきている。Wikipedia とはウィキメディア財団が運営するフリーのオンライン百科事典であり、Wiki システムを導入している。Wikipedia 記事はユーザの手によって編集されるが、その際記事間のリンクは編集者が指定しないと作成されない。これらが自動的に作成されないのは、２記事間が意味的に関係していないのにリンクが作成されるのを防ぐためであるが、逆に意味的に関係しているのに見落とされて作成されないリンクも存在する。これを本研究ではミッシングリンクと呼ぶ。このミッシングリンクの問題が解決されれば、閲覧者が関連情報の提示により更なる理解を深められたり、記事編集者が関連情報を踏まえた上で記事を書いたりすることが出来る。そこで本研究では Wikipedia 記事におけるミッシングリンクを自動的に発見する手法を提案する。

## 2．従来手法

Adafre ら[3]は、Wikipedia 記事におけるミッシングリンクを発見する手法を提案している。

Wikipedia における記事間のリンクは、記事間に意味的なつながりがあると認められる場合に張られることが多い。また、用語の多義性の問題や編集するユーザの書き方の違いなどの問題もあるため、記事中に現れる単語と合致する記事があるからといってそれら全てにリンクを張るのは得策ではない。そこで、ミッシングリンクを探す対象となる記事と関連した記事を探索し、それらのリンクをそのトピックに関連するリンクとし、それらと対象となる記事について比較して不足しているリンクを補うという手法を提案した。

関連記事の探索の方法は以下の通りである（図１）。まず対象記事に対してリンクを張っている記事を探索する。次にそれらが対象記事以外にリンクを張っている記事を探索し、これらを対象記事に対する関連記事候補とする。対象記事と関連記事候補はそれぞれ経由する記事からリンクされているが、このような対象記事と関連記事候補の関係を参照共起関係という。関連記事候補がどのくらい関連しているかを判断する指標については Cocitation[4]を用いる。これはシードページであるミッシングリンクを探す対象記事と参照共起する回数を関連度とする指標である。この関連度が高い上位ｎ件を選択し、関連記事と判断する。

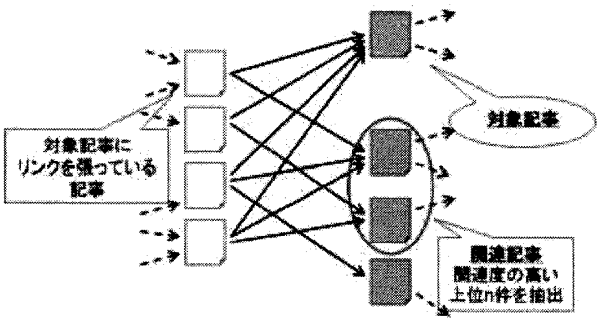


図１　リンク元を辿る関連記事の探索

対象記事に関連した記事が探索できたら、次にミッシングリンクを提案する（図２）。関連記事ではリンクが張ってあったアンカーテキストに対し対象記事で張っていなければ、それをミッシングリンクとして提案する。

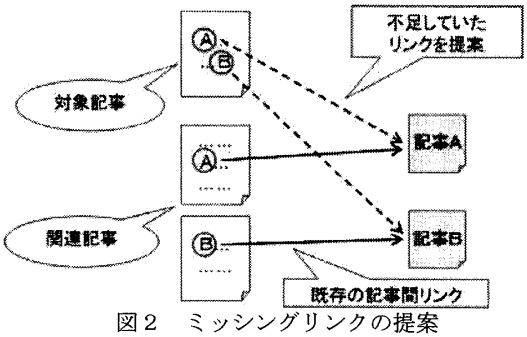


図２　ミッシングリンクの提案

以上が従来のミッシングリンクの発見手法であるが、これには２つ問題点がある。一つは関連記事を探索する際に対象記事のリンク元しか考慮してない点である。新しく作成されたばかりの記事など対象記事の被リンク数が少ない場合、関連度が分散してしまい、ページ遷移の間に話題が変化するトピックドリフトが起きているような記事が関連記事として見つかる可能性が高い。もう一つはリンクの重みを均等に計算している点である。たとえ Cocitation による関連度が同じだとしても、途中リンクを経由している記事が関連記事へ多くリンクを出しているか、少ないリンクしか出していないかでその記事の価値が大きく変わり、それによりリンクの価値も変わってくる。よって経由する記事の価値によりリンクに重みを付けることにより、関連していると判断する基準をさらに高めることで更なる精度向上につながる可能性がある。

## 3．提案手法

本研究では従来手法を改良し、被リンク数の少ない記事に対応し、経由する記事の価値によりリンクに重み付けをして関連度を算出することで関連記事探索の精度を向上させた上でミッシングリンクを発見する手法を提案

"Automatic Discovery of Missing Links in Wikipedia"
†1 Takahito Nakagawa, Graduate School of SIE, University of Tsukuba.
†2 Kazutaka Furuse, Graduate School of SIE, University of Tsukuba.
†3 Hanxiong Chen, Graduate School of SIE, University of Tsukuba.

する。関連記事からミッシングリンクを発見する手法は従来通りである。

まず被リンク数の少ない記事にも対応できるよう、リンク元だけでなくリンク先からも関連記事を探索することにする。対象記事がリンクしている記事を探索し、更にその記事へとリンクしている記事を探索する（図３）。これにより、対象記事が同じような記事をリンクしている記事を探索することができるため、これらを関連記事として扱うことにする。

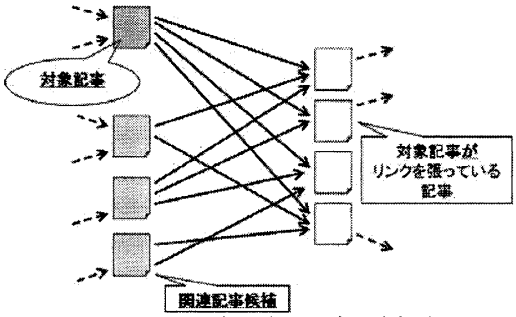


図３　リンク先を辿る関連記事探索

次にリンクの重み付けについては、HITS[5]の考えを用いる。HITS は Web グラフからオーソリティとハブという２つの性質を抽出するアルゴリズムである。関連したページへの情報連絡網の役割を持つハブと、ハブからリンクされることによりトピックに関して権威を持つオーソリティの２つの性質を数値で表す。ページ p のハブスコア H(p)及びオーソリティスコア A(p) は以下の式により算出される。

$$H(p)=\sum_{p\to q}A(q) \qquad (1)$$

$$A(p)=\sum_{q\to p}H(q) \qquad (2)$$

オーソリティスコアが高ければ、そのトピックに関しては権威のあるページであることが伺える。一方でハブスコアが高ければ、そのトピックに関しては優秀なリンク集として機能するページであることが伺える。優秀なリンク集から張られているリンクや、権威のあるページへのリンクの価値は高く評価できる。

そこで提案手法では、ハブスコアの高い記事から出ているリンクと、オーソリティスコアの高い記事を指しているリンクの重みを高く設定し、Cocitation では比較できなかったリンクの価値を考慮した関連度により関連記事をランキングする。リンク元を辿って関連記事を得る方では、途中で経由する記事と関連記事との間についてハブスコアの高い経由記事からのリンクに高く重みを付ける。一方でリンク先を辿って関連記事を得る方では、途中で経由する記事と関連記事との間についてオーソリティスコアの高い経由記事へのリンクに高く重みを付ける。

対象記事 $p_i$ と関連記事候補 $p_j$ について、$p_i$ および $p_j$ へのリンクを有するページ集合を $X(p_i, p_j)$、$p_i$ および $p_j$ へからのリンクを有するページ集合を $Y(p_i, p_j)$とすると、$p_i$ と $p_j$ の関連度 $S(p_i,p_j)$は以下の式により算出される。$\alpha$は重み付け定数である。

$$S(p_i,p_j)=(1-\alpha)\sum_{p\in X(p_i,p_j)}H(p)+\alpha\sum_{p\in Y(p_i,p_j)}A(p) \qquad (3)$$

この式は、ハブスコアの総和やオーソリティスコアの総和を算出する際にリンクの数などで正規化を行う必要があるかどうかについて議論し、必要であれば改善する余地があると考えられる。

## ４．予備実験

対象記事のリンク数が少ない場合に従来手法が非効率的であることを確かめるため、従来手法について関連記事を探索し、その内容を評価した。

対象記事として、被リンク数が多い「Direct3D」と、被リンク数が４と比較的少ない「新井淑子_(音楽家) 」の２つの日本語版 Wikipedia 記事を採り上げた。その結果、記事「Direct3D」では関連度上位に高い関連度で関連性のある記事が多く占めていた。一方で記事「新井淑子_(音楽家)」では、関連度のごく上位には関連性が高いと考えられる記事では占めているものの、その関連度は一律低く、記事数も少なかった。またそのすぐ下には一見関連性の無い記事が多く抽出された。また、関連度上位と下位では関連度にあまり差が出ていないため、関連していない記事と関連している記事とが全体的に混在する形となった。

以上の結果から、対象記事のリンク数が少ないと効率良く関連記事を探索できないことが分かった。

## ５．まとめと今後の課題

本研究では、Wikipedia 記事におけるミッシングリンクの自動発見手法について、従来手法の問題点を改善した新しい手法を提案した。

今後は提案手法の有効性を確かめるべく評価実験を計画している。実験方法としては、ミッシングリンクを発見する対象となる記事のリンクを一つ隠しておき、従来手法と提案手法のそれぞれで関連記事の探索を行い、それらを用いてミッシングリンクの提案を行う。隠したリンクが見つかったかどうかについて、その再現率を算出し比較することにより提案手法の有効性を確かめたいと考えている。提案した関連度算出の式等、まだ議論すべき部分があるため、実験の結果から提案手法の改良を行っていきたいと考えている。

## 参考文献

[1] Wikipedia 英語版. http://en.wikipedia.org/wiki/
[2] Wikipedia 日本語版. http://ja.wikipedia.org/wiki/
[3] S. F. Adafre and M. de Rijke, “Discovering missing links in Wikipedia”, Proceedings of the 3rd International Workshop on Link Discovery at KDD05, pp.90-97, 2005.
[4] Jeffrey Dean, et al, “Finding Related Pages in the World Wide Web”, Computer Networks, 31, pp.1467-1479, 1999.
[5] J. Kleinberg, “Authoritative sources in a hyper-linked environment”, Proc. 9th ACM SIAM Symposium in Discrete Algorithms, pp.668-677, 1998.