インターネットからの情報獲得と統合化

Information Gathering and Integration with the Internet

武田英明
Hideaki Takeda
奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

1 はじめに

近年 WWW(World Wide Web) に代表されるインターネットの情報サービスが普及するにつれ、莫大な情報からどのように自分の必要な情報を獲得していくかが問題になってきた。この問題を解くひとつの鍵は検索技術であるが、もうひとつの鍵は情報の知的統合である。本稿では特に後者に絞り、情報空間からいかに情報を獲得するか、そしてそれを統合するかという問題（ネットワークを用いた知的情報統合 [16]）について議論する。

2 ネットワークを用いた知的情報統合

ネットワークを用いた知的情報統合には、これまで独立した分野として研究されてきた様々な分野が関係してくる。そこで、図 1はネットワークを用いた知的情報統合に関わる概念を摸式的に示したものである。主な概念としては、ネットワークからどう情報を集めるかといういう情報収集、その情報をどう利用者の目的に合わせて変化させるかという情報統合、そういった情報のやりとりを同実現するかという情報流通の３つがある。

まず、ネットワークからの情報収集 (information gathering) としては、情報検索 (information retrieval)、情報フィルタリング (information filtering)[13]、ブラウジング (browsing) の方法が挙げられる。情報検索はユーザが欲する情報の仕様が明確でありかつ情報源が現にアクセス可能な時に有効である。情報フィルタリングでは、情報源が動的に変わる場合など現時点ではアクセスできない場合に有効である。ブラウジングは欲する情報が明確でない時有効な方法である。ブラウジングをどう知的にする試みは Web Watcher[1] や Letizia[6] などがある。

情報の統合としては、情報分類 (information classificatoin)、情報抽出 (information extraction)、情報組織化 (information organizing) などが挙げられる。分類は収集しきた情報を適当なカテゴリーに分けることであり、クラスタリングなどの方法がこれまで行なわれてきた。抽出とは収集してきた情報から必要な情報を抜き出すことであり自然言語理解などに関連する。構造化とはさらにそれらの情報の関係づけを行なうことであり、発想支援などとの関わりがある。情報抽出の例としては、佐藤らの電子ニュースのダイジェスト自動作成 [15]、後述の IICA などがある。情報組織化の例としては CM-2[7] がある。

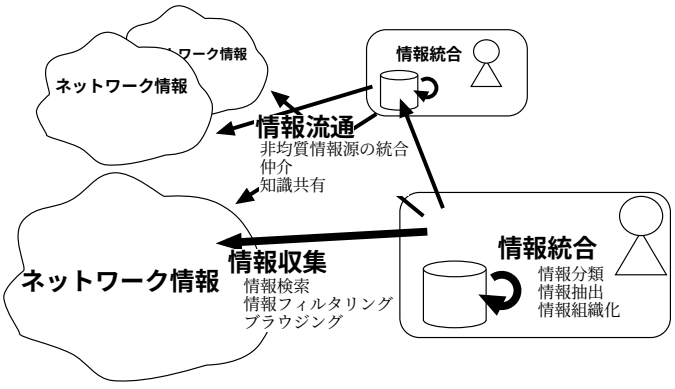

図 1: ネットワークを利用した知的情報統合の枠組み

ネットワークを用いた知的情報統合では、これら情報統合が情報収集と独立に行なわれるのではなく、お互いに影響しあって行なわれるところに特色がある。また、情報統合を自分が行なうのではなく、他のところで行なったものを利用する場合も考えられる。

また情報流通の問題もある。これは情報源や情報利用者が多数かつ異なる場合、どのようにして情報源と情報利用者の関係を適切につくることができるかという問題である。具体的には非均質情報源の統合、や仲介、知識共有などがテーマとなる。

3 技術的背景

前節で述べたようにネットワークを用いた知的情報統合には様々な研究分野が関連してくる。そのなかで、特に鍵となる技術・研究は以下の点である。

1. 柔軟なテキスト検索技術
2. 頑強な自然言語処理
3. 知識体系
4. ユーザのモデル

2. はテキストを理解するという意味ではこれまでの自然言語処理であるが、ここで必要なのは詳細な内容理解ではなく、どういった内容が書いていあるかといった概略の理解である。このためより浅いが頑強な自然言語処理、たとえば統計的な処理やテンプレートを併用する方法 (FAQ Finder[2])、共起関係を利用した方法 (例えば METIS[10][11]) などが用いられている。

3. は広範な情報の関連性を知るための背景知識である。例えば語彙体系や概念体系の利用が考えられる。汎用な概念体系を目指したものとしては IICA のオントロジーがある。また、語彙体系としては WordNet[12][5] を利用しているものもある (FAQ Finder)。

4. はユーザの利用に沿った情報収集や情報統合ができるようシステムにとって必要である。ユーザの嗜好の学習としては、 Web Watcher[1] などの学習するブラウザや一定の自律性をもちユーザの代行を行なうインタフェース・エージェント（例えば Maes の Learning Interface Agents[8][9]）などが挙げられる。

```
(define-pclass (温泉 ((has-one 温泉の名前)
                      (is-a 訪問地)
                      (has-some 風呂の種類)
                      (has-some 泉質)
                      (has-some 効能))))
(define-concept (効能 (is 傷病 with (or "効能>" "効果"
                                         "効く"))))
(define-concept (傷病 (or "+症>" "+傷>" "+病>")))
...
```

図 2: 属性概念の記述例

URL	温泉の名前	最寄り駅	アクセス方法	風呂の種類	泉質
akase-spa-j.html	"赤瀬温泉"		"バス"		"炭酸鉄泉"
hinagu-spa-j.html	"日奈久温泉"	"JR八代駅"	"JR日奈久駅下車"		"食塩泉" "単純"
kanaketa-spa-j.html	"金桁温泉"	"JR三角駅"	"バス"		"炭酸鉄泉"
tsurugiyama-spa-j.html	"鶴木山温泉"	"JR佐敷駅"			"単純"
tsuruyu-spa-j.html	"鶴湯温泉"		"徒歩"		"単純"
yoshio-spa-j.html	"吉尾温泉"	"JR吉尾駅"	"徒歩"		"単純"
yunoko-spa-j.html	"児温泉"	"JR水俣駅"	"バス"	"沖合いの湯"	"重曹泉"
yunotsuru-spa-j.html	"鶴温泉"	"JR水俣駅"	"バス"		"単純"
yunoura-spa-j.html	"湯浦温泉"	"JR湯浦駅"	"徒歩"		"単純"

図 3: 情報抽出例

4 システムの実際

ここではIICA (Intelligent Information Collector and Analyzer) を例にとり、インターネットからの情報統合の実際をみていくことにする。

IICA はWWW から情報収集・分類・抽出を行なうシステムである [3][4]。このシステムの特徴はオントロジーと呼ばれる概念体系を情報の収集・分類・抽出の 3 つの段階全てを通して利用することである。ここでの概念とはその特徴を示すキーワードの集合とその属性を文章上での出現パターンで記述したものである。後者は情報抽出のとく用いられる。概念間は重みをつけた有向のリンクで結ばれたネットワークを作る。

情報収集　情報の収集ではオントロジー上で指定した概念に関するページを最初に指定したページからリンクでつながっているページを順次見ていくことで収集する。この際、次のリンクをみにいくかどうかは指定した概念およびその概念にオントロジー上で近傍にある概念に関するキーワードを含むかどうかで決定する。

情報分類　IICA における情報の分類は情報収集フェーズで収集したページをオントロジー上の概念に割り付けることである。ここでは各ページに対して空間ベクトル法 [14] によるキーワードの出現頻度を要素とする特徴ベクトルを計算し、これと各概念に対する特徴ベクトルとの類似度を計算することで、ページを分類する。特徴ベクトルはその時にその概念に分類されるページの特徴ベクトルに平均値である。

情報抽出　各概念には前述のようにその属性的概念を文章中に現れるパターンを用いて記述しているので、これを用いて各ページから必要とする情報を抜き出す。属性概念の記述の例を図 2に示す。この方法を用いて同じ概念に分類されたページから表を作ることができる (図 3参照)。

5 まとめ

インターネットの情報を取り扱うシステムには従来のシステムに比べ、異質かつ乱雑な情報の特質を活かすような方法が必要である。本稿ではそのための技術として柔軟なテキスト検索技術、頑強な自然言語処理、知識体系、ユーザのモデル、を挙げた。また、システムの実例として、 WWW から情報収集・分類・抽出を行なうシステム IICA の紹介を行なった。

参考文献

[1] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell. Webwatcher: A learning apprentice for the world wide web. In *1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, pp. 6–12, 1995.

[2] R. Burke, K. Hammond, and J. Kozlovsky. Knowledge-based information retrieval from semi-structured text. In *1995 AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, pp. 15–19, 1995.

[3] 岩爪, 白神, 畑谷, 武田, 西田. テキストからの情報抽出・統合化法の提案と知的情報収集・分析システム IICA の実験的評価. 第 7 回データ工学ワークショップ, 1996.

[4] M. Iwazume, H. Takeda, and T. Nishida. Ontology-based information capturing from the internet. In *Proceegins of the fourth International Conference on the International Society of Knowledge Organization*, 1996. (To appear).

[5] C. S. Laboratory. WordNet. `http://www.cogsci.princeton.edu/~wn/`.

[6] H. Lieberman. Letizia: An agent that assists web browsing. In *Proceedings of IJCAI-95*, pp. 924–929, 1995.

[7] 前田, 西田. 知識メディアシステム CM-2 とそのユーザインタフェース. 第 11 回ヒューマンインタフェースシンポジウム論文集, pp. 49–54, 1995.

[8] P. Maes. Agents that reduce work and information overload. *Communications of the ACM*, 37(7), 1994.

[9] P. Maes and R. Kozierok. Learning interface agents. In *Proceedings of AAAI-93*, pp. 459–465, 1993.

[10] 松尾, 武田, 西田. KP 化による論文内容の効果的提示方法とその応用. ヒューマン・インターフェス・シンポジウム, pp. 581–588, 1995.

[11] T. Matsuo and T. Nishida. Intelligent support for construction and exploration of advanced technological information space. In *Proceegins of the fourth International Conference on the International Society of Knowledge Organization*, 1996. (To appear).

[12] G. A. Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[13] D. Oard. Information filtering resources, March 1996. `http://www.enee.umd.edu/medlab/filter/`.

[14] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1983.

[15] 佐藤, 佐藤, 篠田. 電子ニュースのダイジェスト自動作成. 情報処理学会論文誌, 36(10):2371–2379, 1995.

[16] 武田. ネットワークを利用した知的情報統合. 人工知能学会誌, 10(5), 1996. (掲載予定).