

WWW テキストの自動要約と KWIC インデックスの作成

清田 陽司 黒橋 禎夫

京都大学大学院情報学研究科 知能情報学専攻

〒 606-8501 京都市左京区吉田本町

kiyota@pine.kuee.kyoto-u.ac.jp kuro@i.kyoto-u.ac.jp

本論文では、自動要約によって WWW テキストへの KWIC インデックスを作成するシステムについて述べる。本システムは、WWW テキストを収集する WWW ロボット、TF-IDF 法にもとづいて WWW テキストから重要文を抽出するモジュール、さらに重要文を単文や句の形に圧縮する文要約モジュールで構成される。文要約モジュールは、KNP による構文解析結果を用いて文を単文や句の単位で分割し、TF-IDF 法にもとづいて重要な部分のみを取り出す。各モジュールを評価したところ、重要文抽出モジュールでは 74.5% の精度が得られ、文要約モジュールでは 200 文中 177 文で満足できる要約が得られた。

Automatic Summarization of WWW Texts and its Application to a WWW KWIC Index

Youji Kiyota Sadao Kurohashi

Graduate School of Informatics, Kyoto University

Yoshidahonmachi, Sakyo, KYOTO 606-8501 JAPAN

kiyota@pine.kuee.kyoto-u.ac.jp kuro@i.kyoto-u.ac.jp

This paper presents a system which creates a KWIC index of WWW texts by automatic summarization. The system consists of three modules: a spider for WWW, an important sentence extractor from texts, and a sentence summarizer. The most effective module is the last one which employs a robust and fairly accurate parser KNP. It segments an input sentence into phrases or simple sentences and assembles a summary. The accuracy of the important sentence extractor was 74.5% and that of the sentence summarizer was 88.5%.

する高品質なインデックスを作成することが可能となる。

文よりも小さな単位での要約に関する研究としては、[亀田 95]、[山本ほか 95]、[三上ほか 99] などがある。これらの研究では、簡単な構文解析を行い、文末の用言を残し、修飾や例示などの部分を固定的な規則で削除することによって要約を行う。しかし、実際には必ずしも文末が重要であるとは限らず、より柔軟に文中から重要な部分を取り出す必要がある。そのためには、埋め込み文、並列構造などの文の正確な構造を把握する必要がある。

本論文では、頑健で高精度な日本語構文解析システム KNP[黒橋 98] を利用し、KNP によって得られる種々の言語的情報と、TF-IDF 法による語の重要度の尺度を組み合わせるにより、より柔軟に文の要約を行う方法を提案する。

2 WWW テキスト重要文抽出モジュール

WWW テキスト重要文抽出モジュールは、Salton が提案した TF-IDF 法 [Salton 89] を主な手掛かりとし、これに WWW テキストの記述言語である HTML の構造を組み合わせて用いる方法で、重要文の抽出を行う。

2.1 WWW テキストからの文抽出

まず、WWW テキストから文の抽出を以下のルールによって行う。

- 句点「.」、疑問符「?」、感嘆符「!」は文区切りとして扱う。ただし、括弧 (鍵括弧や丸括弧など) 内にあるものは文区切りとはしない。
- 以下の HTML タグを文区切りとみなす。
HR, P, BR, TITLE, HEAD, BODY, H1 ~ H6, CENTER, DIV, BLOCKQUOTE, PRE, XMP, LISTING, PLAINTEXT, UL, OL, DIR, MENU, LI, DL, DT, DD, TABLE, CAPTION, TR, TH, TD, THEAD, TBODY, TFOOT
- 整形済みテキストを表すタグ (PRE, XMP, LISTING, PLAINTEXT) で囲まれる部分については、改行コードを文の区切りとみなす。
- フォーム (FORM) や Java Script (SCRIPT) など で囲まれる部分は、テキストとして扱わない。

表 1: HTML 属性にもとづくキーワードの重みづけ

属性	タグ	重みづけ
タイトル	TITLE	16 倍
キーワード	META(keywords)	32 倍
説明	META(description)	10 倍
見出し	Hn ($n = 1 \sim 6$)	(9 - n) 倍
センタリング	CENTER, DIV(center)	1.2 倍
箇条書き	UL, OL, DIR, MENU	0.8 倍
テーブル	TABLE	0.8 倍

2.2 TF-IDF 法によるキーワードの重みづけ

抽出された各文を形態素解析してキーワードを抽出し、各キーワードの重要度スコアを TF-IDF 法により計算する。

まず、抽出された各文について、日本語形態素解析システム JUMAN[黒橋ほか 99] を用いて形態素解析を行い、品詞分類が普通名詞 / サ変名詞 / 固有名詞 / 地名 / 人名 / 組織名 / カタカナ / アルファベットである語をキーワードとして抽出する。

つづいて、TF-IDF 法を用いて、収集した WWW テキスト集合に含まれる各テキストに対する全てのキーワードの重要度スコアを計算する。テキスト集合に含まれるテキスト数を N とすると、テキスト D_i に対するキーワード k_j の重要度スコア $w(i, j)$ は、以下のように計算される。

$$w(i, j) = TF_{ij} \cdot \log \left(\frac{N}{N_j} \right) \quad (1)$$

ただし、 TF_{ij} はキーワード k_j のテキスト D_i における出現回数、 N_j はキーワード k_j の出現するドキュメント数である。

2.3 WWW テキストの構造を考慮したキーワードの重みづけ

HTML によって示される WWW テキストの構造は、語の重要度の大きな手掛かりとなる。多くの場合、タイトルや見出しに含まれる語はテキストにとっての重要語であり、一方、リストやテーブルに含まれる語はあまり重要とはいえない。よって、タイトルや見出しに含まれるキーワードに高いスコアを、リストやテーブルに含まれるキーワードに低いスコアを与えることにする。本論文では、キーワードの出現する位置にもとづいて、表 1 の重みづけを行う。

これは、上の式 (1) において TF_{ij} の数え方を修

正することを意味する。例えば、タイトルと見出し(H1)部分に「金融」というキーワードが各1回現れる場合は、これを24(=16+8)回分の出現としてカウントする。

2.4 重要文の選択

2.2節、2.3節の方法で得られたキーワードの重要度スコアを用いて、各文の重要度スコアを計算する。文の重要度スコアは、文に含まれるキーワードの重要度スコアの和を、キーワード数の n 乗で正規化したものとする。すなわち、テキスト D_i 中の文 S_t が延べ m_t 個のキーワードを含み、その重要度スコアが $w(i, 1), \dots, w(i, m_t)$ であるとすると、文 S_t の重要度スコア $I(i, t)$ は、

$$I(i, t) = \frac{\sum_{j=1}^{m_t} w(i, j)}{(m_t)^n} \quad (2)$$

となる(4章の実験では $n = 0.5$)。

また、WWWテキストにおいては、重要文が存在する範囲は比較的限られており、以下のことが経験的にいえる。

- 重要文は、本文の先頭数文の範囲に含まれることが多い。
 - 極端に短い文が重要文であることは少ない。
 - 箇条書き部分に重要文が含まれることは少ない。
- そこで、重要文を抽出する範囲を以下のように絞込む。

箇条書き属性をもつ文と c 文字未満の文を除く、
先頭より a 文

以上の範囲で式(2)のスコアが最も大きい文を重要文として抽出する(4章の実験では $c = 15, a = 6$)。

3 言語知識にもとづく文要約モジュール

文要約モジュールは、以下の処理を行うことによって文の圧縮を行う。

1. KNPによる構文解析結果、要約文の制限文字数、文が含まれるテキストにおける各キーワードの重要度スコアを入力として受け取る。
2. いくつかの言語的規則によって、構文木を意味的なまとまりに分割する。このまとまりを以下ではパートと呼ぶ。
3. 分割された各パートの重要度スコアを、TF-IDF法にもとづいて計算する。

4. 重要度スコアの最も高いパートを要約の核とする。
5. 残りのパートの中から、重要度スコアの高い順に、制限文字数を超えない範囲で核に連結していく。
6. 連結できる残りのパートがなくなったら、結果を要約として出力する。

3.1 構文木の分割

KNPによって構文解析された文を言語的に意味を持つ小さなパート、すなわち単文または名詞句に分割する。このような単位で分割された各パートは、構文木上で隣接している限りどのように結合しても文として意味をなすと考えられる。

KNPが各文節に与える属性を用い、以下に述べる規則に従って構文木を分割する。

連用節の分割 連用節(連用形の用言に係る部分構文木)と主節を分割する。ただし、複合辞をなす用言、すなわち「(～に)対して」「(～と)して」などは、主節との強い結合を示すと考えられるので切り離さない。

連体節の分割 連体節(連体形の用言に係る部分構文木)と主節を分割する。このとき、連体修飾を受ける体言は主節のパートと連体節のパートの双方に含める。ただし、連体節の係り先が形式名詞(「こと」「もの」「とき」など)、または外の関係にある名詞句(「問題」「方針」など)である場合は、修飾を受ける体言だけでなく連体節全体を主節のパートに含める。例えば、「本社を移転することに致しました」という文は、「本社を移転する」と、「本社を移転することに致しました」の2つのパートになる。

デ格の分割 格助詞「で」によって示されるデ格には場所、道具、材料、原因などの用法があるが、いずれの場合でも係り先の用言との結び付きは比較的弱いので、主節から切り離す。

副詞、接続詞の分割 副詞、接続詞は、省略しても大方の意味は失われないので、切り離して扱う。

時間を表す名詞句の分割 「本日」「昨年」「この度」などの時間を表す名詞句は、一般には省略可能である。ただし、「本日は晴天なり」のように必須格と考えられる格に入っているものは省略できない。よって、時間を示す名詞句のうち、無格、ノ

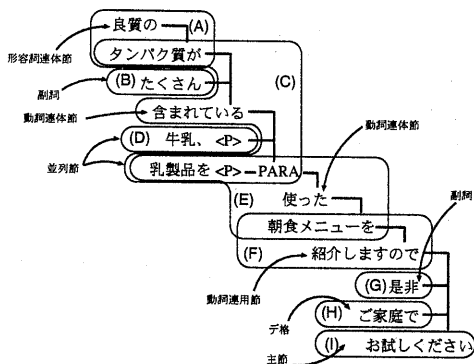


図 3: 構文木の分割の例

格，カラ格，マデ格，隣接に係るものは切り離す。
並列要素の分割 並列をなす部分は，多くの場合一部の並列要素を省略しても意味をなす。よって，まず各要素の重要度スコアを計算したのち，最も重要度スコアの大きな要素以外を切り離す (3.2節で詳述)。

この分割規則の適用例を図 3 に示す。分割された各パートが意味を持った小さなまとまりになっていることがわかる。

なお，分割された各パートを個別にみた場合，末尾の助詞などは省略可能である。よって，各パートの末尾に存在する以下の形態素の連続は**末尾省略可能表現**として扱う。

- すべての助詞
- 助動詞「のだ」の全ての活用形 (「ので」など)
- 判定詞「だ」の全ての活用形 (「です」など)
- 句点「。」，読点「、」

3.2 各パートの重要度スコア計算

分割された各パートの重要度スコアを，重要文抽出と同様にキーワードの重要度スコアより計算する。すなわち，重要と思われる語を多く含むパートを，重要なパートとみなす。

まず，文に含まれる各キーワードの重要度スコアを，2 章の方法で計算する。図 2 のシステム構成図で示した通り，単語頻度 (TF) は重要文が含まれるテキストにおける値，文書頻度の逆数 (IDF) は WWW テキスト集合における値を用いる。

こうして得られた各キーワードの重要度スコアを用

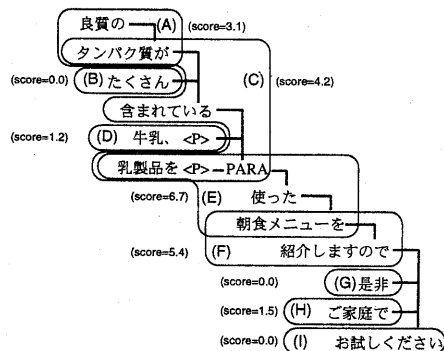


図 4: 各パートの重要度スコア計算の例

いて，各パートの重要度スコアを計算する。

- パートが含むキーワードの重要度スコアを加算し，キーワード数の n 乗で正規化したものを基本スコアとする。すなわち，パート i が延べ $k(i)$ 個のキーワードを含み，それぞれの重要度スコアを $W(i, 1), \dots, W(i, k(i))$ とすると，基本スコア $S(i)$ は，

$$S(i) = \frac{\sum_{j=1}^{k(i)} W(i, j)}{k(i)^n} \quad (3)$$

となる (4 章の実験では $n = 0.5$)。

- 文末のパートの重要度スコアは，基本スコアを m 倍したものとする (4 章の実験では $m = 2.0$)。これは，日本語においては文末のパートは文の主節であり，大抵の場合は重要と考えられるからである。
- 並列要素のうち，主節から切り離されて独立したパートとなっているものは，基本スコアを l 倍したものとする (4 章の実験では $l = 0.5$)。
- 副詞，接続詞，時間を表す名詞句の重要度スコアは 0 とする。すなわち，これらのパートは，文字数に余裕があるときにのみ要約に含められることになる。

図 3 の例について，各パートの重要度スコアを計算した例を図 4 に示す。

ここで，並列要素に関する重要度スコア計算とパート分割について補足説明を行う。この文では，2 つの名詞句「牛乳」「乳製品 (を)」が並列句をなす。それぞれを一つのパートとしてみた場合の基本スコアは，「牛乳」が 2.4，「乳製品 (を)」が 2.8 となる。よって，後者の「乳製品 (を)」がこの並列句を

代表する名詞句として扱われ、主節「使った朝食メニュー(を)」に併合される。結果として、2つのパート「乳製品を使った朝食メニュー(を)」「牛乳」ができる。それぞれのスコアは、前者が6.7、後者は $2.4 \times l = 2.4 \times 0.5 = 1.2$ となる。

3.3 制限文字数内での要約の生成

全てのパートの中で、最も大きな重要度スコアを持つパートを要約の核として選択する。もし、この核の文字数が制限文字数以上であれば、これをそのまま要約として出力する。

こうして選択された要約の核に、合計の文字数が制限文字数を超えない範囲で、スコアの順に一つずつ結合していく。

1. 現在選択されている要約パート集合(最初は核のみ)と係り受け関係をもつパート(隣接パート)の中で、最も重要度スコアの大きなパートを選ぶ。残りのパートが存在しなければ、現在の要約パート集合を要約として出力する
2. 選ばれたパートを含めた要約の文字数を調べる。
3. もし文字数が制限文字数以内であれば、このパートを要約パート集合に追加する。制限文字数を超えるならば、このパートを破棄する。
4. 1に戻る。

ただし、現在の要約パート集合に「～ため」「～ので」「～について」に係るパートは、主節(現在の要約パート集合)との強い結び付きを表すと考えられるので、重要度スコアを基本スコアの s 倍とする(4章の実験では $s = 1.5$)。

図4についてアルゴリズムを適用した場合のステップを以下に示す(制限文字数を25文字とした)。

1. 最も重要度スコアの大きいパート(E)「乳製品を使った朝食メニュー(を)」(14文字)を核として選ぶ。
2. 次に重要度スコアの大きい隣接パート(F)を選択する。これを核に結合すると、要約は「乳製品を使った朝食メニューを紹介します(ので)」(19文字)となり、制限の範囲に収まる。
3. 次に重要度スコアの大きい隣接パート(C)を選択する。これを結合すると、要約は「タンパク質が含まれている乳製品を使った朝食メニューを紹介しま

す(ので)」(31文字)となるが、制限文字数を超えるのでパート(C)は放棄される。

4. 次に重要度スコアの大きい隣接パート(D)を選択する。要約の文字数は22文字となる。
5. 残りのいずれのパートも追加すれば制限文字数を超えるので、要約を終了する。

この結果、要約として「牛乳、乳製品を使った朝食メニューを紹介します」が出力される。

4 実験と考察

本システムにより、Yahoo! Japan¹、京都大学ホームページ²を起点に日本語で記述されたWWWテキストを収集し、KWIC インデックスを作成した。現在、約70万個(約4Gbytes)のWWWテキストへのインデックスとなっている。

この一部のWWWテキストを用いて、重要文抽出モジュール、文要約モジュールの評価を行った。

4.1 重要文抽出精度の評価実験

重要文抽出モジュールの精度を、テストテキスト集合を用いて評価する実験を行い、結果を考察した。

ロボットにより収集したWWWテキスト集合より、ランダムにテキストを選び出した。ただし、リンクやリストのみからなるテキスト、日記などのテーマ非限定のテキストは重要文を抽出する意味がないものとして人手により除外し、結果として200個のテキストをテスト集合とした。そして、それぞれのテキストについて重要と考えられる文(テキストのインデックスとして利用することが適当と考えられる文)を重要文としてマークし、重要文抽出の実験の正解とした。1テキストあたりの平均重要文数は1.69であった。

そして、2.2節で述べたキーワードの出現頻度のみを用いた重要文抽出の方法(出現頻度のみ)と、2.3節で述べたWWWテキストの構造も利用した重要文抽出の方法(テキスト構造利用)について、それぞれ評価実験を行った。

評価は、各テキストについて、システムによって最大の重要度スコアを得た文が、人手によって重要文とマークされたものである場合を正解とした。評価の結果を表2に示す。表2に示すように、テキスト構造を

¹ <http://www.yahoo.co.jp/>

² <http://www.kyoto-u.ac.jp/>

表 2: 重要文抽出モジュールの評価結果

	テキスト構造利用	出現頻度のみ
正解	149 (74.5%)	82 (41.0%)
不正解	51	118
合計	200	200

利用する重要文抽出の方法は、キーワードの出現頻度のみを用いる方法と比較して著しく良い精度を達成しており、正解率は74.5%であった。

重要文抽出では、2.3節で述べたように抽出範囲の絞り込みを行っている。この範囲内にマークされた重要文が含まれていたのは200テキスト中185テキスト、すなわちこの絞り込みの精度は92.5%であり、この絞り込みは非常に有効であった。

4.2 文要約精度の評価実験

文要約モジュールの精度を、テスト文集合を用いて評価する実験を行い、結果を考察した。

前節と同様に、ロボットにより収集したHTMLテキスト集合からランダムにテキストを選び、それぞれの最重要文を手手で選んだ。このとき、最重要文が50文字未満であるものは除外し、結果として200個のテスト重要文を選択した。テスト重要文の長さの平均は80.5文字であった。

要約の制限文字数を25文字または45文字とし、評価実験を行った。各キーワードの重要度スコアの計算には、ロボットにより収集した約70万個のテキスト集合を用いた。

生成された要約に対し、人手によって以下の3段階で評価を行った。

- ： 原文の大意が適切に要約されている
- △： 重要な情報の一部が欠落しているか、文としておかしいところがあるが、原文の大意は理解できる
- ×： 原文の大意が失われている

評価の結果を表3に示す。○と△を合わせた割合は、制限文字数を25文字とした場合で88.5%であり、45文字とした場合は93.5%であった。

評価が×であった例について原因を調べたところ、主な要因は以下の3つであった。

1. KNPが文節間の係り受けの解析を誤った場合、不適切な要約を出力することがある。

表 3: 文要約の評価結果

	制限文字数	
	25文字	45文字
○	109 (54)	153 (18)
△	68 (13)	34 (3)
×	23 (3)	13 (0)
合計	200 (70)	200 (21)

(括弧内の数値は、要約の核の文字数が制限を超えている文の数)

2. TF-IDFによる単純な重みづけがうまくいかないため、あまり情報を含まないパートに最も大きなスコアが与えられた場合、要約結果はあまり良くない。
3. 長い名詞句を含む大きなパートが要約の核として選ばれる場合がある。このような場合、文字数制限のために他のパートを追加することができず、しばしば重要な情報が失われる。

評価が×であった例がどの要因にあてはまるかを調べたところ、1が約50%、2が約30%、3が約20%であった。すなわち、要約の精度を下げる要因のおよそ半分は、構文解析の誤りによるものであった。

また、評価が△であった文要約において目立った問題は、文内、文章内での照応や省略が考慮されていないことである。これを補うことで、より良い要約を生成することができると考えられる。

図5に、テスト文と、制限文字数を25文字とした場合の要約結果の例を示す。

例1は、文末(主節)のパート「暴力団離脱者社会復帰協議会が開かれた」が要約の核として選ばれ、良い要約が生成されている例である。これは、日本語においては主節が一般に重要であることを示している。これは例2も同様である。一方、例3における主節「ページも増えてきているように思います」は、相対的に重要な情報ではない。文要約モジュールは、「カヌーで」を要約の核として選び、結果として良い要約が生成されている。

例4の要約では、「ナスダック・ジャパンに共同出資する」主体が何であるかという重要な情報が失われている。これは、要約の核が埋め込み文であり、この節の主語が構文的には主節「調印しました」に依存しているからである。もし省略されている主語「ソフトバンクおよび全米証券業協会」が推定できれば、より適切な要約「ソフトバンク(など)がナスダック・ジャ

	テスト文	要約結果と評価
1	暴力団組織から離れたたいという組員に助言したり、就職先を紹介する <u>暴力団離脱者社会復帰支援協議会の総会が29日午前、県警本部で開かれた</u> (65 文字)	暴力団離脱者社会復帰支援協議会の総会が開かれた (24 文字) ○
2	今回は中級編のまとめということで、今まで作ってきた関数をまとめて「 <u>画像加工</u> 」アプリケーションを作成してみます (54 文字)	今回は「画像加工」アプリケーションを作成してみます (25 文字) ○
3	インターネットも市民権を持ってきて、カヌーで出かけられる川の情報を掲載したページ もだいたい増えてきているように思います (58 文字)	カヌーで出かけられる川の情報を掲載したページ (22 文字) ○
4	ソフトバンク及び全米証券業協会は1999年6月15日、日本におけるまったく新しい電子証券市場を開設する新会社、 <u>ナスダック・ジャパンに共同出資する契約</u> に調印しました (81 文字)	ナスダック・ジャパンに共同出資する契約 (19 文字) △
5	県総務部は26日までに、各知事部局と県職労に対し、知事部局の組織定数計4570人の人員配置について、187人削減し、61人増員させるなどとした1999年度の <u>組織定数配置案を提示</u> (88 文字)	県総務部は組織定数配置案を提示 (15 文字) ×
6	平成11年6月現在における市町村が設置するごみ焼却施設のコンピュータ西暦2000年 <u>問題への対応状況</u> 等について調査を行った (60 文字)	問題への対応状況等について調査を行った (19 文字) ×

図 5: 文要約の例

バンに共同出資する契約」を生成することができる。

一方、例 5、6 は×と判断された要約例である。例 5 では、「各知事部局と県職労に対し」は要約に含める必要があると考えられるが、KNP はこの係り先が文末ではなく「～増員させる」であると誤って解析したため、要約に含めることができなかった。例 6 は、名詞句「コンピュータ西暦 2000 年問題」の一部「西暦 2000 年」が時間要素であると誤って解析されたことが原因である。

5 おわりに

本論文では、自動要約によって WWW テキストへの KWIC インデックスを作成するシステムについて述べた。本システムの要点は、言語的情報と単語の TF-IDF によって文をさらに圧縮することにある。これにより、WWW テキストへのコンパクトなインデックスを作成することが可能となった。

現在、http://www-lab25.kuee.kyoto-u.ac.jp/service/www_kwic/ にて KWIC インデックスを参照することができる。今後、KWIC インデックスの有用性を大規模な実験で検証する予定である。

謝辞

本研究を進めるにあたり適切な助言および貴重なご意見をいただきました京都大学の河原達也助教授に深

く御礼申し上げます。

参考文献

- [奥村ほか 98] 奥村 学, 難波 英嗣: テキスト自動要約技術の現状と課題, 北陸先端科学技術大学院大学情報科学研究科 技術報告 IS-RR-98-0010I (1998)
- [亀田 95] 亀田 雅之: 日本語文書読解支援系 QJR の検討, 情報処理学会研究報告, Vol. 95, No. 110, pp. 57-64 (1995)
- [山本ほか 95] 山本 和英, 増山 繁, 内藤 昭三: 文章内構造を複合的に利用した論説文要約システム GREEN, 自然言語処理 Vol. 2, No. 1, pp. 39-55 (1995)
- [三上ほか 99] 三上 真, 増山 繁, 中川 聖一: ニュース番組における字幕生成のための文内短縮による要約, 自然言語処理 Vol. 6, No. 6, pp. 65-81 (1999)
- [黒橋 98] 黒橋 禎夫: 日本語構文解析システム KNP version 2.0 b6 使用説明書, 京都大学大学院 情報学研究科 (1998)
- [Salton 89] Gerard Salton: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley Publishing Company (1989)
- [黒橋ほか 99] 黒橋 禎夫, 長尾 真: 日本語形態素解析システム JUMAN version 3.62 使用説明書, 京都大学大学院 情報学研究科 (1999)