Unstructured Information Management Projects at IBM Tokyo Research Laboratory

Koichi Takeda, Hideo Watanabe, Naohiko Uramoto, Hiroshi Nomiyama, Hirofumi Matsuzawa, Tetsuya Nasukawa, Tohru Nagano, Akiko Murakami, Hironori Takeuchi, Hiroshi Kanayama, Mei Kobayashi, Masaki Aono*, Akihiro Inokuchi, Michael E. Houle**

Content

1. Introduction

Unstructured Information Management(UIM) research and development has always been one of the central topics for Asian countries for more than three decades, since the localization of computer systems and content in native languages was critical for leveraging leading-edge technologies into every aspect of the academic and business activities in each country. In many Asian countries, the first challenge was to develop an input method for their native languages. This laid the foundations for morphological analysis, machine-readable dictionaries, and the notion of "disambiguation" that accompanies almost all technical issues in UIM. Machine Translation (MT) was the next big thing in the 1980's, and promoted lexical, syntactic, and semantic analysis methods. MT systems for translating computer manuals were studied by many IBM research groups because of the importance to IBM's business. During this period, it became clear that corpus-based and machine-learning approaches needed to be incorporated to implement scalable and easily customizable machine translation systems, since handcrafted systems were so time-consuming and expensive. It also turned out that many theoretical and constraint-based grammar formalisms were not powerful enough to identify only one analysis for an input sentence in many practical cases, and stochastic methods(or ad hoc scoring) had to be employed to obtain the most likelyanalysis among alternative candidates. In the late 90s, MT systems for translating between English and Japanese were commoditized -- although with gisting translation(i.e., primarily for understanding, and not for publication) quality --in Japan due to the overwhelming demand for Web page translation. Millions of copies of MT systems were shipped with PCs in Japan.

The UIM research has been shifting its focus toward text mining since the late 90s. With billions of Web pages available for search via the

* Toyohashi University of Technology, Japan

** National Institute of Informatics, Japan

Internet, text mining would be a key differentiator for search and analytic solutions and services including question answering, market intelligence, and monitoring/alerting. Information retrieval(IR) will be enhanced significantly with text mining(in particular, information extraction) technology. Many enterprise business innovations such as knowledge management, customer relationship management, and content manage-ment, would also call for text mining from vast numbers of corporate documents. Information Integration[7, 11] (II) intersects with UIM in this area. There are many specialized databases, application-oriented data files(e.g., spreadsheet data), and documents that constitute a broad and distributed collection of information sources. In particular, life sciences research involves many databases for genomics and proteomics information, sequence data, and millions of biomedical journal abstracts(such as MEDLINE). UIM and II are expected to provide means for navigating, cross-referencing, and obtaining patterns and hidden facts from a wide range of information sources. In the sections below, we will describe the past and current UIM research activities at IBM Tokyo Research Laboratory (TRL) and give perspectives of upcoming trend in UIM and related research areas.

2. Machine Translation

The goals of the MT project are to facilitate globalization of enterprise contents and to support collaborative communication among international employees as globalization, contents management, and instant messaging/collaboration are the growing interests of corporate customers. Our earlier efforts were made for building rule-based MT systems for translating between English and Japanese languages[32, 15]. These systems employed the transferapproach, where syntactic structures in source and target languages are converted by a rule-based procedure. A number of approaches pre-transferring source syntactic structures into canonical forms[32], interactive human pre-editing of text[15], and constraint dependency grammar[16] for describing Japanese syntactic structures were proposed and implemented, but the difference inquality between human and MT translations was still significant. TRL also started a joint project with Carnegie-Mellon University for a knowledge-based MT(KBMT) project[6]. Domain knowledge and ontology were built for im-plementing the interlingual representations(i.e., language-independent meaning representations) of source and target sentences. The interlingua-based MT appeared to be more elegant than the conventional, transfer-based MT approach, since the former approach requires just one analyzer andone generator for each source and target language, while the latter approach requires the source-to-target language transfer component for each language pair in addition to the analyzer/generator. The KBMT approach was further explored at TRL[28] between 1989 and 1995 by incorporating many practical and theoretical aspects of MT, including bi-directional grammars, conceptual paraphrasing[29], and contextual disambiguation[20].

These MT systems had a limited success, but it was not favorably accepted by the professional human translators of computer manuals in general, largely due to the fact that post-editing work(modifying the MT output and making the final translations) was tiresome, and MT could save only a modest portion of the entire translation workflow. Even nowadays, a successful MT deployment can often be found in environments, where a large number of users are non-professional translators, and they appreciate turn-around time, ease-of-use, and gisting translations as good enough for their needs and time saving. By late 1990s, we also realized the workload of building and updating knowledge sources was so high that we needed an alternative approach to meet emerging demand for Internet Web page translation in Japan, which aims to translate a chaotic collection of text in virtually unrestricted domains and linguistic styles. We therefore employ pattern- and corpus-based approach toward MT[33, 30, 34]. A pattern is a pair of context-free grammar(CFG) rules, each for source and target language, with a cost, fixed number of binary features, and a headconstraint (described below) for non-terminal symbols. The purpose of using patterns is to augment the syntax-directed translation scheme[2] and synchronous grammars[26] for incorporating lexicalized rules and syntactic constraints such as subject-verb agreement more easily. It was also inspired by the notion of translation by analogy [19, 25], which appeared to be quite promising for accumulating translation knowledge for a diverse collection of Web pages. For example, the pattern (for simplicity, binary features are not shown)

be:V:1 year:NP:2 old -> VP:1
VP:1 <- avoir:V:1 an:NP:2

describes that the verb(V) "be", followed by a noun phrase(NP) "year"(presumably modified by a numeral expression) and the adjective "old"in English, makes a verb phrase(VP), whose features are given by the verb. The pattern also associates this left-hand side CFG rule with the right-hand side French CFG rule such that the verb "avoir" takes a noun phrase "an" as an object. The stems "be", "year", "avoir", and "an"are constraining the syntactic head of the constituents, and are called the head constraints. Patterns are recursively combined to give a pair of syntax trees for source and target sentences. In the above example, the English noun phrase "5 years" should have its French counterpart noun phrase "5 ans"based on a separate pattern for a numeral modifying a plural noun. Note that the above pattern can translate English phrases "is almost two years old" and "are exactly 300 years old", and the resulting verb phrase can be combined with another pattern for associating it with a noun phrase as subject to constitute a sentence. Wecould quickly build a large collection of patterns some are purely phrasalpatterns with no lexicalized constituents, and most others are lexicalized patterns for each class of syntactic constituents. The collection of patterns showed a steady, incrementalimprovement of the translation quality against a sample evaluation set of English sentences and Web pages as the size of the collection grew. It was rather remarkable, since the conventional transfer-approach suffered non-monotonic changes in translation quality when the collection of transfer rules is modified. It was also confirmed that novice users can easily define patterns for customizing the domain-specific translations.

While patterns are extensively used for English-to-Japanese translation to directly associate two surface structures, they are defined as transferrules between two internal representations for Japanese-to-English trans-lation. This is because of accommodating rather free word-ordering sentences in Japanese. A semi-automated approach has been developed for extracting lexicalized transfer rules from bilingual corpora, and the collected transfer rules are complied into a decision tree for converting a source representation into a target one. We are also proposing a modified BLEU method[23] for automating the MT quality evaluation of agglutinative languages such as Japanese[12]. Another interesting extension of our MT research is to improve the translation language pairs by incorporating an annotated hub language(e.g., English)[13] instead of an artificial interlingua, so that we can achieve translation between any two languages by providing a component for translating between each language and the hub language. Since the annotated hub language representations generated by the source-to-hub language translator include lexical and structural information as annotations, there would not be degradation of translation quality that would normally be caused by syntactic ambiguities of hug language representations (without annotations) when they are fed into the hub-to-target language translator. This is becoming one of key issues in the MT research, as the Web pages are getting more and more globalized and written in many different languages.

3. Text Mining

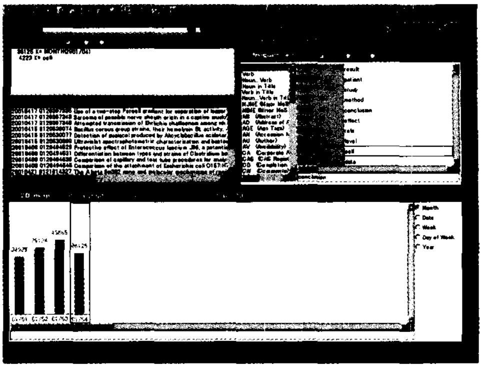
The notion of text mining seems to have emerged circa 1995, when Feldman and Dagan proposed an approach for knowledge discovery from textual databases[5]. It has been conceived that text mining could mean clustering a large number of documents into subcollections of documents with similar contents, characterizing such subcollections for visual data/document exploration[18], or routing a new document into one of predefined categories. As pointed out by Hearst[8], text mining is very different from information retrieval, since the former emphasizes on discovery of unknown nuggets(such as facts, patterns, and regularities), while the latter aims to find the documents most relevant to a user's expressed information request. More significant potential of text mining, however, turns out to be in information extraction. Customer Relationship Management(CRM), in particular, requires fine-grained textual analysis for problem identification, trend discovery, and reputation mining from millions of customer inquiries and claims to a contact center and open discussions in the Web forums and bulletin boards. Information extraction allows us to identify named entitiessuch as including product names, organizations, places, and their associations(e.g., subject-verb and verb-object relationships). Once such information extraction technology is established, wide ranges of UIM applications and solutions could be built upon the derived information. In particular, the vector-space modelof documents, where each document is represented by a vector of weighed extracted information, allows us to calculate similarity between documents. This is a basis for similarity search, categorization, clustering, and many machine learning applications of text mining.

3.1 IBM TAKMI

Through our experiences with IBM PC Help Centers and external customers since 1998, it appears the following is a list of key ingredients for text mining for most of use cases.

1. term dictionary and taxonomy for(possibly hierarchical) term categorization
2. named entity extraction and shallow parsing for phrasal chunking
3. dependency analysis for relation extraction
4. stochastic mining algorithms for singularity, pattern, and association discovery

Term dictionary and taxonomy are major knowledge sources for customization of text mining. For example, IBM PC Help Centers made use ofhardware and software terms and taxonomy for their problem identification and trend analysis. Shallow parsing and relation extraction are critical for understanding underlying issues from customer call logs, since words and phrases are often too abstract to represent factual information (that is, four terms "LAN adaptor", "installation", "need", and "support" are not quite informative as two relations "support--LAN adaptor" and "need--installation").



(Figure 3.1) Overview of the IBM TAKMI Text Mining System

We have also observed that modal, mood, and intention information(usually represented by auxiliary verbs, adverbs, and collocational expressions) is extremely important to text mining. Unlike information retrieval, where only content words are used to index documents, such information has to be appropriately annotated together with content words. For example, three expressions "modem is broken", "modem isn't broken", and "Is modem broken?" are indistinguishable with respect to content words included in the expressions, but they represent totally different customer intention complaint, neutral statement, and inquiry that has to be identified for making appropriate analysis and actions in CRM.

In (Figure 3.1), we show our text mining system, called IBM TAKMI(Text Analysis and Knowledge Mining)[21]. TAKMI users can either enter keywords or select a term in the category view to define a subcollection of documents to be mined. Publication dates(year, month, etc.) in the chronological distribution view can also be used to slice the subcollection. Six mining methods with associated visualization(called views) are in-corporated into TAKMI 2D map view, delta graph view, topic extraction view, and topic 2D view in addition to two viewsshown in the (Figure 3.1)(see Nasukawa and Nagano[21] for more details). By switching these views, the users can find characteristic terms and relations in a selected category, salient topics, and correlations between terms in the specified subcollection of documents. The views not only provide mining results, but allow users to navigate and further narrow down the subcollection to get more relevant documents. Instead of reading individual documents, the users can acquire contextual and summary information mined from the sub-collection even if they do not intend to search for specific documents. The design of TAKMI views are motivated by the notions of information visualization[27] and hierarchical nature of textual information. Even a single customer contact log can include multiple fragments of information such as complaints and inquiry, which implies assignment of single log categoryis difficult and sometime misleading. To capture every key fragment of information necessary for knowledge discovery, text mining would have to produce a huge amount of extracted information and annotations, and effective means of exploring (overview, zoom, and details on demand[27]) the extracted information had to be implemented. In addition to set-oriented abstraction(counting, sorting, and grouping) of terms, text mining can provide linguistically-motivated abstraction of extracted information. For example, there are several levels of textual information:(the numbers in parentheses denote the number of contact logs including the information. These samples are shown just for illustrative purposes)

1. SPAM (105)
2. SPAM mail (68)
3. SPAM filter (24)
4. hadSPAM mail [complaint] (11)
5. needSPAM filter? [inquiry] (6)

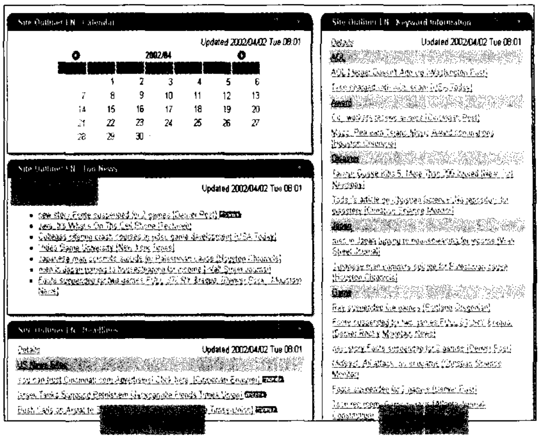
The term "SPAM" can be used to identify all the contact logs potentially relevant to the topic about "SPAM". The two compound words "SPAM mail" and "SPAM filter" can then be used to focus on more specific topics. Finally, the relations "hadSPAM mail" and "needSPAM filter?" should be descriptive enough to learn major SPAMissues. The(absolute) frequency of terms above may not be adequate to learn about minor, but important issues. In this case, another type of frequency(called relative frequency) can be used. The relative frequency of a term T is the ratio of document frequency(the number of documents) including the term T in the current subcollection, divided by the document frequency of the term T in the entire document collection. By ranking the terms and relations in terms of relative frequency, we can often identify salient terms and relations in a specific subcollection of documents.

TAKMI has been widely accepted by customers in cross-industrial sectors mainly for CRM enhancement. It can work with Japanese and English documents. Our named entity extraction method for Japanese text has been integrated into DB2 Information Integrator for Content. Recently, TAKMI has been applied to MEDLINE a collection of 12 million biomedical journal citations for life science text mining, and first deployed by a biotech company in Japan since 2002.

One of the latest developments at TRL's text mining is sentiment analysis and reputation mining[22]. It captures emotive and sentiment expressions including a given topic word, and summarizes favorability and unfavorability of the topic. This technology looks very promising for market intelligence and information risk management such as fraud, rumors, and false accusation of a company or its products in the Internet.

3.2 Site Outlining

When it comes to mining from the World Wide Web, many of the Web pages behave more like a broadcasting program. They keep on updating their content, and conventional text mining methods are not adequate for such Web mining, since those methods assume only static information sources. It has also been observed in the Internet that finding fresh information(and feeding it) from the information flood has a heavy demand. These observations were made earlier by Douglis et al. in their paper[3]. The "push" technologies were then introduced around 1996to deliver fresh information to the Internet users, and the delivery service achieved more than one million subscribers around the world in 1997. These ideas now seem to be revived in WebLog


Site Outliner EN - Calendar
Updated 2002/04/02 Tue 08:01
2002/04
1 2 3 4 5 6
7 8 9 10 11 12 13
14 15 16 17 18 19 20
21 22 23 24 25 26 27
28 29 30
Site Outliner EN - Top News
Updated 2002/04/02 Tue 08:01
Site Outliner EN - Headlines
Details
Updated 2002/04/02 Tue 08:01
US News Sites
Site Outliner EN - Keyword Information
Details
Updated 2002/04/02 Tue 08:01
AOL
Award

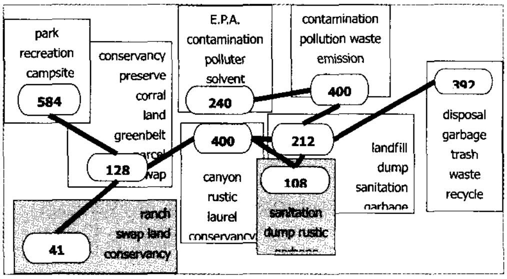
(Figure 3.2) Site Outlining Portlets

(e.g., Robot Wisdom log in http://www.robotwisdom.com/).

Site outlining [31] has been developed for handling such dynamic information sources. It consists of crawling the specified Web pages regularly, archiving them as consecutivesnapshots, information extraction from the crawled Web pages, identification of new information -- anchors(annotated links) and paragraphs, and mining methods over the newly obtained information. It can therefore create a digest of more than hundreds of Web pages by showing (a)what's new, (b)what's hot, and (c)top news. "What's new" is a collection of new, annotated links, "what's hot" refers to a set of named entities(person and organization names) with intensive appearances in the latest snapshots, and "top news"are the collections of annotated links that share almost the same set of keywords and published in multiple Web pages in other words, the news that are covered by many Web pages. For each of the category above, we can associate a score(or importance) with the news, calculated based on the freshness, difference between occurrence frequencies of named entities in current and previous period(usually a week), the number of keywords in the filtering list, and the number of Web sites that publish the news. It is also possible that by archivingthe metadata of the snapshots of each Web page, we can reconstruct a "weekly magazine" or "monthly magazine "views for any(even daily updated) Web page by arranging the newly added articles by day or week. This longer-term characterizationof Web contents would facilitate more reliable and accurate Web site categorization/clustering as well as identification of salient topics in the Web site. Unlike ordinary Web search and ranking based on single snapshot of each Web page, a revised search and ranking method based on multiple snapshots of Web pages could result in a considerably different hit list and providean alternative ranking scheme. Each Web page can also be classified into four types: active (frequently updated), periodical(periodically updated) inactive(rarely updated), and archival (with no update since a long time ago).

In (Figure 3.2), we show site outlining implemented as portlets for top news, what's new, and filtered articles based on user-defined keywords. Site outlining was also implemented as IBM Japan's premium consumer software, mySiteOutliner, for its paid-membership sub-scribers, and was bundled with IBM PCs in Japan. The information extraction and mining techniques for site outlining could be incorporated into a very large system such as WebFountain [4] to provide business intelligence for billions of Web pages.

3.3 Vector-Space Technologies



(Figure 3.3) Overlapping Clusters Obtained from LA Times News Articles

The vector-space model (VSM)[24] represents documents as vectors of weighted terms, and it has been widely used for information retrieval (IR) and text mining. At TRL, VSM is incorporated into similarity search(or search by example) technology and local clustering(i.e., clusters with overlapping neighborhood collections of documents)[9]. Dimensional reduction of huge, but very sparse VSM is also explored [14].

Our local clustering method is characterized by four notions:

(1) pairwise distance measure dist(d_1, d_2) between two documents d1 and d2 in the document database D,

(2) k-nearest neighbor set NN(D, d, k) of any given document d in D,

(3) confidence conf(C_1, C_2) between two clusters C_1 and C_2, based on their size and the number of common neighbors (intersection) such that conf(C_1, C_2) = $|C_1 \cap C_2| / |C1|$, and

(4) self-confidence sconf(C) of a given cluster C,

based on the average confidence of the cluster C with its neighboring clusters(of the same size) such that $sconf(C) = \Sigma conf(C, NN(D,d,|C|)/|C|$ for all d in C.

By varying the size k(within a limited range) of nearest neighbors for each document d, we can observe a significant drop of self-confidence, which indicates the boundary of the best candidate cluster. By successively constructing clusters for each subcollection of the documents $\{S_0, S_1, , S_i, , S\}$, where $Si = |S|/2^i$ for $0 \langle i \langle \log |S|$, and by pruning duplicates and associating inter-cluster links between two clusters with high confidence, we can obtain the cluster graph as illustrated in the (Figure 3.3). Overlapping clusters appear to be essential for knowledge discovery from real-world documents, since each document often consists of multiple topics, and partitioningclusters would often result in several major clusters and very fragmented minor clusters. Major clusters usually turnout to represent known topics, while many of novel topics can be found in minor clusters. In (Figure 3.3), we show a partial collection of overlapping clusters obtained from 127,738 LA Times news articles prepared for the 6th Text Retrieval Conference(TREC). The squares represent clusters. The terms in each square are representative keywords of the cluster, and the size of each cluster is shown by the circled number. The solid line between two clusters indicates that the connected clusters are strongly associated. There are two minor clusters shown in the figure one cluster with 41 documents and another cluster with 108 documents. The former cluster(concerned with a specific land management controversy) represents an interesting subtopic of another cluster, while the latter cluster(concerned with the use of canyons as dumping grounds) heavily overlaps with two major clusters, and may not be discovered by partitioning clustering methods. In addition to its capability for knowledge discovery, our clustering method is quite scalable. It has successfully processed more than one million MEDLINE abstracts into 7,260 clusters.

4. Information Integration

As described in the previous sections, the UIM technologies are now capable of turning unstructured text data into a rich collection of objects terms, names entities, intention and sentiment expressions, relations, sentential structures, and so on. Information integration(II) can incorporate textual information -- not just as searchable fields, but as a collection of fine-grained textual objects -- into the database world. Many of database-orientedtechnologies can also be applied to textual objects combined with database attributes. Extraction of frequently asked questions (FAQs) is one of them. FAQs are defined as frequent syntactic structures (common subtrees) found in the collection of inquiry sentences, and the Apriori algorithm [1] for finding item sets can be extended to structural patterns[17]. Similarly, in life sciences the chemical structures can also be analyzed for associating particular common chemical substructures with a given property(such as toxicity) described in a database. Our graph mining method[10] has shown to be the most efficient algorithm for such a complex and tricky (becauseof graph isomorphism) domain. As the DNA microarray analysis becomes a routine work for identifying clusters of genes with particular functions, II can extend the analysis by making sense of clusters with insights extracted from biomedical documents. Since textual descriptions are full of associations between biomedical entities, UIM and II should work as bread and butter to provide vital information for pathway and protein-protein interaction analysis.

We have been applying these UIM and II technologies to many life sciences resources (MEDLINE, US life sciences patents, GenBank and other databases) for a broad range of knowledge discovery, which would benefit drug discovery as well as clinical solutions. It is also important to note that many of these textual resources are now available in XML. In addition to the analytics aspects, XML native·store and querying functions would become more and more important in this area.

Another promising area of II might be business intelligence. For example, if a company would like to monitor its competitors in the market, there are plenty of information sources such as patents, news feeds, Web pages, and US Securities and Exchange Commission(SEC) filing to be analyzed and integrated to make a rich intelligence report. UIM and II technologies can semi-automate much of the analysis and reporting work, with a broader coverage and deeper analysis.

5. Conclusion

In this paper, we described the UIM research activities at TRL. When we look back, it turns out machine translation projects laid the solid foundation of UIM technologies and possibly many new technologies to come. Text mining connected the UIM research in information extraction and analysis with many industrial applications. While machine translation concentrated on a technology for mapping representations in two languages and preserving original information content, text mining leads us to a whole new technology that handles potentially unbounded amount of information derived from the original information content. It has also been recognized that in the structured information management community, objective information(i.e., facts and data) is arguably the most valuable and appreciated in the data processing. In the UIM community, however, text mining provides an alternative view of information by incorporating subjective in-formation to express wide ranges of properties associated with objective information certainty, belief, attitude, respect, usefulness, and so on. These properties would then be used to classify, filter, and score objective information for knowledge management and collaboration support.

The role of textual information will become more and more important in many industrial applications and solutions, since a vast amount of heterogeneous resources must be integrated for supporting a streamline, on-demand chain of business processes. Textual information would work as a nice glue to correlate such diverse resources, possibly with the help of common taxonomy and ontology. In particular, text mining and II for life sciences are developing more rapidly than ever before and will continue to develop even faster, since so many resources are now available for worldwide research use, and new methodologies are quickly deployed for these shared resources.

IBM TAKMI is a trademark of International Business Machines Corporation in the United States, other countries, or both.

References

[1] Agrawal, R., and Srikant, R., "Fast algorithms for mining association rules", Proc. 19th Int. Conf. on Very Large Data Bases (VLDB), pp. 487-499, 1994.

[2] Aho A. V., and Ullman, J. D., The Theory of Parsing, Translation, and Compiling, Volume 1: Parsing, Prentice-Hall, Englewood Cliffs, New Jersey, 1972

[3] Douglis, F., Ball, T., Chen Y.-F., and Koutsofios, E., "WebGUIDE: Querying and Navigating Changes in Web Repositories", Proc. of 5th Intl. World-Wide Web Conf. (WWWC), pp.1335-1344, 1996

[4] Edwards, J., McCurley, K., and Tomlin, J., "An Adaptive Model for Optimizing Performance of an Incremental Web Crawler", Proc. of 10th Intl. WWWC, pp.106-113, 2001

[5] Feldman, R., and Dagan, I., "Knowledge discovery in textual databases (KDT)", Proc. of 1st Int. Conf. on Knowledge Discovery and Data Mining (KDD), pp.112-117, 1995.

[6] Goodman, K. and Nirenburg, S. (eds), The KBMT Project: A case study in knowledge-based machine translation, ISBN-1-55860-129-5, Morgan Kaufmann Publishers, San Mateo, CA, 1991.

[7] Haas, L. M., Schwarz, P. M., Kodali, P., Kotlar, E., Rice, J. E., and Swope, W. C., "DiscoveryLink: A system for integrated access to life sciences data sources",IBM Sys. Journal, Vol.40, No.2, pp.489-511, 2001

[8] Hearst, M. "Untangling Text Data Mining" (invited paper), Proc. of the 37th Annual Meeting of Associations for Computational Linguistics (ACL), 1999

[9] Houle, M. E., "Navigating Massive Data Sets via Local Clustering", Proc. of 9th Int. Conf. on Knowledge Discovery and Data Mining (KDD), pp. 547-552, 2003

[10] Inokuchi, A., Washio, T., and Motoda, H., "Complete Mining of Frequent Patterns from Graphs: Mining Graph Data", Machine Learning, Vol.50, No.3, pp.321-354, 2003

[11] Jhingran, A. D., Mattos, N., and Pirahesh, H., "Information integration: A research agenda", IBM Sys. Journal, Vol.41, No.4, pp.555-562, 2002

[12] Kanayama, H., "Paraphrasing Rules for Automatic Evaluation of Translation into Japanese", Proc. of 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003), 2003

[13] Kanayama, H. and Watanabe, H., "Multilingual Translation via Annotated Hub

Language", Proc. of MT Summit IX, 2003

[14] Kobayashi, M., Aono, M., and Houle, M., "Mining Overlapping Major and Minor Clusters in Massive Databases", Proc. 5th Int. Congress on Industrial and Applied Mathematics (ICIAM), 2003

[15] Maruyama, H., Watanabe, H. and Ogino, S., "An Interactive Japanese Parser for Machine Translation", Proc. of 13th Intl. Conf. on Computational Linguistics (COLING), pp.257-262, 1990

[16] Maruyama, H., "Structural disambiguation with constraint propagation", In The Proc. of the 28th Annual Meeting of ACL, pp. 31-38, 1990

[17] Matsuzawa, H., and Fukuda, T., "Mining Structured Association Patterns from Database", Proc. of 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp.233-244, 2000

[18] Morohashi, M., Takeda, K., Nomiyama, H., and Maruyama, H., "Information Outlining - Filling the Gap between Visualization and Navigation in Digital Libraries", Proc. Int. Symp. on Digital Libraries, pp. 151-158 1995.

[19] Nagao, M. "A framework of a mechanical translation between Japanese and English by analogy principle", in "Artificial and Human Intelligence" (Eds: Elithorn, A. and Banerji, R.), North-Holland, Amsterdam,, pp. 173-180, 1984

[20] Nasukawa, T., "Robust Parsing Based on Discourse Information: Completing Partial Parses of Ill-Formed Sentences on the Basis of Discourse Information", Proc. of 33rd Annual Meeting of ACL, pp.39-46, 1995

[21] Nasukawa, T., and Nagano, T., "Text analysis and knowledge mining system", IBM Sys. Journal, Vol.40, No.4, pp.967-985, 2001

[22] Nasukawa, T., and Yee, J., "Sentiment Analysis: Capturing Favorability Using Natural Language Processing", to appear in Proc. of 2nd Int Conf. on Knowledge Capture (KCAP), 2003

[23] Papieni, K, Roukos, S., Ward, T., Zhu, W.-J., "BLEU: A Method for Automatic Evaluation of Machine Translation", Proc. of the 40th Annual Meeting of ACL, pp.311-318, 2002

[24] Salton, G., Wong, A., and Yang, C. S., "A vector space model for automatic indexing", Communications of the ACM, Vol.18, No.11, pp.613-620, 1975.

[25] Sato, S., and Nagao, M. "Toward memory-base translation", Proc.of the 13th Intl. Conf. on COLING, pp.247-252, 1990

[26] Shieber, S. M., and Schabes, Y., "Synchronous Tree-Adjoining Grammars", Proc. of 13th Int. Conf. on COLING, pp.253-258, 1990

[27] Shneiderman, B., Designing the User Interface: Strategies for Effective Human-Computer Interaction, 2nd Edition, ISBN:0-201-57286-9, Addison-Wesley, Reading, MA (1992).

[28] Takeda, K., Uramoto, N., Nasukawa, T., and Tsutsumi, T., "Shalt2: A Symmetric Machine Translation System with Conceptual Transfer", Proc. of 14th COLING,

pp.1034-1038, 1992.

[29] Takeda, K., "Tricolor DAGs for Machine Translation", Proc. of 32nd Annual Meeting of ACL, pp. 226-233, 1994

[30] Takeda, K., "Pattern-Based Context-Free Grammars for Machine Translation", Proc. of 34th Annual Meeting of ACL, pp.144-151, 1996

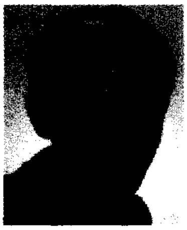
[31] Takeda, K., and Nomiyama, H., "Site Outlining", Proc. of 3rd ACM Digital Libraries Conf., pp.309-310, 1998

[32] Tsutsumi, T., "A Prototype English-Japanese Machine Translation System", In "Natural Language Processing: The PLNLP Approach", (eds: Jensen, K., Heidorn, G., and Richardson, S.), ISBNKluwer Academic Publishers, Boston, 1993

[33] Watanabe, H., "A Similarity-Driven Transfer System", Proc. of 14th Intl. Conf. on COLING, pp.770-776, 1992

[34] Watanabe H., and Takeda, K., "A Pattern-based Machine Translation System Extended by Example-based Processing", Proc. of Int. Conf. on COLING-ACL 2000, pp.1369-1373, 2000

저자약력



Koichi Takeda

Koichi Takeda joined the IBM Tokyo Research Laboratory in 1983 after receiving his ME degree in Information Science from Kyoto University. His research interests include machine translation, text analytics, and information retrieval. He was a visiting researcher at the Carnegie-Mellon University during 1987-1989 for carrying out the joint research project on Knowledge-based Machine Translation