Intelligent Tickers: An Information Integration Scheme for Active Information Gathering

Yasuhiko Kitamura
kitamura@info.eng.osaka-cu.ac.jp
Department of Information and Communication Engineering
Graduate School of Engineering, Osaka City University
3-3-138 Sugimoto, Sumiyoshi-ku, Osaka 558-8585, JAPAN

Abstract.

An active information gathering system efficiently collects information from a number of frequently updating information sources on the Internet, considering the quality and cost of information gathering, to meet demands from human users or active mining systems. In this paper, we summarize functions required for active information gathering systems and show related works. We then propose a system called Intelligent Ticker. Intelligent Ticker consists of multiple information gathering modules and an information integration module. An information gathering module produces Tickers based on the difference between an updated Web page and the original one. The information integration module integrates multiple Tickers by using an ITT (Integration Template for Tickers) to assist the user in his decision making or problem solving.

1 Introduction

The Internet rapidly spreads into our society as one of infrastructures that support our daily life. Among a number of Internet based information services, the WWW (World Wide Web) is most popular and widely used for various purposes such as sharing information among researcher to advance research and development, disseminating sales information for electronic commerce, creating virtual communities that share a common interest, and so on. Considering the vast amount of various information stored on the Web, we may be able to regard the Web as a world wide knowledge base system on the Internet.

The most important feature of the Web is that it is built up in a bottom up manner, which is contrasting with conventional distributed database systems. Once an information provider connects his/her computer to the Internet and starts a Web server, he or she can immediately disseminate information toward the world. The Web can be viewed as a federated system where a huge number of distributed information sources are running autonomously and cooperating with each other without any centralized control mechanism.

On the other hand, from a viewpoint of information users, the Web has a drawback that it is not easy to locate required information in the huge amount of data widely distributed on the Internet. As remedies to deal with this drawback, various search engines have been developed. To a query with input keywords, however, a search engine sometimes just returns thousands of URLs, which often include ones unrelated to the user's request, and the user has to filter the output.

To make the Web more useful, we further continue to study technologies for not only improving outputs of Web search engines, but also developing new systems, which we call active mining systems, that can automatically discover useful information from a huge number of Web information sources by employing various techniques such as machine learning, information agents, information retrieval, database systems, computer human interaction, and so on.

To develop active mining systems, we need to consider the following features of Web information.

- Web information is widely distributed over a huge number of Web sites in the world.
- Web pages are normally described in HTML (Hyper Text Mark-up Language). HTML is suitable for representing the visual structure of Web page, which displays on a Web browser, but not for representing the semantic structure. To deal with this drawback, XML [1], which enables information providers to represent the semantic structure by inserting semantic tags into Web pages, has been standardized. Moreover, research on Semantic Web [2] aims at enabling computers to process the Web information without human interventions based on the XML standardization.
- The amount of Web information increases very rapidly day by day and a large number of Web sites are updated frequently. Especially, ones that deal with stock market or Internet auction are updated almost every minute. Even an active mining system succeeds to discover some information from such sites, if the information is obsolete, it is useless for the user. The active mining system should keep monitoring the Web sites and modify the discovered information depending on the updates of the original sites.

This paper discusses active mining systems that discover useful information for the user through gathering information from a number of Web information sources that may be updated frequently. An active mining system is a data mining system that mainly mines data gathered through the Internet. The activeness of active mining system comes from the dynamics of information sources (updates of information sources) and the user (changes of user's interests or requests), and an active mining system consists of three modules as shown in Figure 1.

- The active information gathering module monitors a number of Web sites, which is dynamically updated, on the Internet and gathers Web pages from them to provide them to the data mining module.
- The data mining module analyzes data gathered by the active information gathering module and discovers information useful to the user.
- The active reaction module plays a role of the user interface. It shows information discovered by the data mining module. By monitoring the user's response to the output, this module notifies changes of the user's interest or request to the data mining module.

The process of active mining is performed by three modules cooperating with each other. This paper focuses the discussion on the active information gathering module.

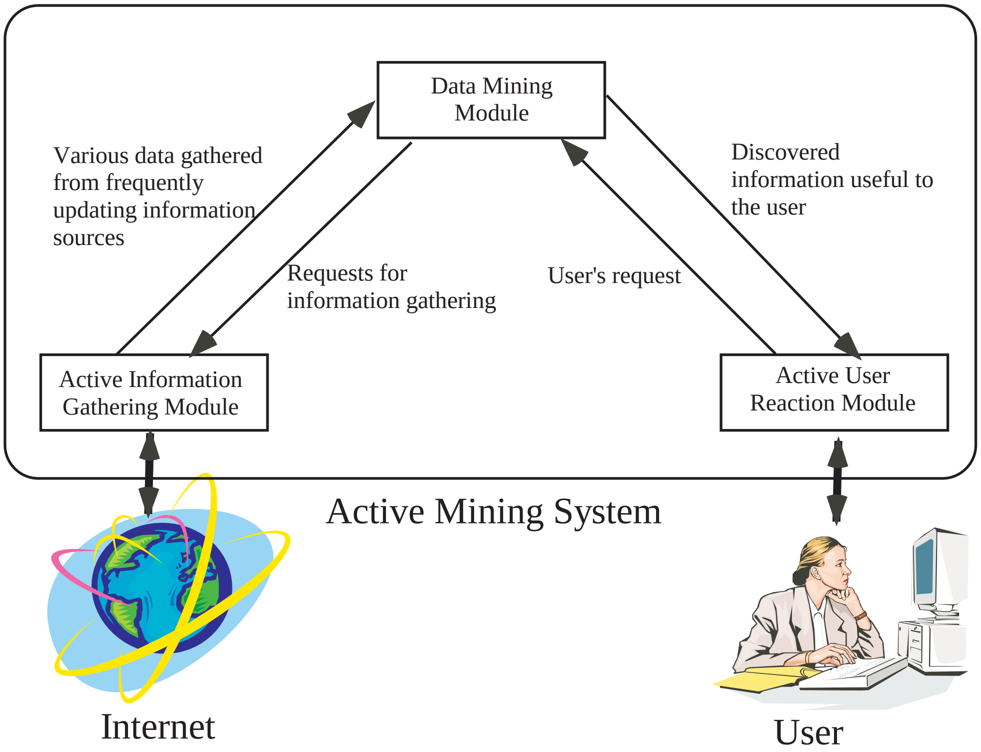

Data Mining Module
Various data gathered from frequently updating information sources
Requests for information gathering
User's request
Discovered information useful to the user
Active Information Gathering Module
Active User Reaction Module
Active Mining System
Internet
User

Figure 1: An Overview of Active Mining System

In Section 2, we discuss functionalities required for the active information gathering module and introduce some related works. In section 3, we propose Intelligent Ticker, which is an active information gathering system based on Ticker, which is a small piece of information extracted based on the difference between the updated Web page and the original one. We conclude this paper in Section 4.

2 Active Information Gathering Systems

Active information gathering system monitors Web information sources that are frequently updated and efficiently gathers information that meets requests given by the user. The functionalities required for active information gathering systems are summarized as follows.

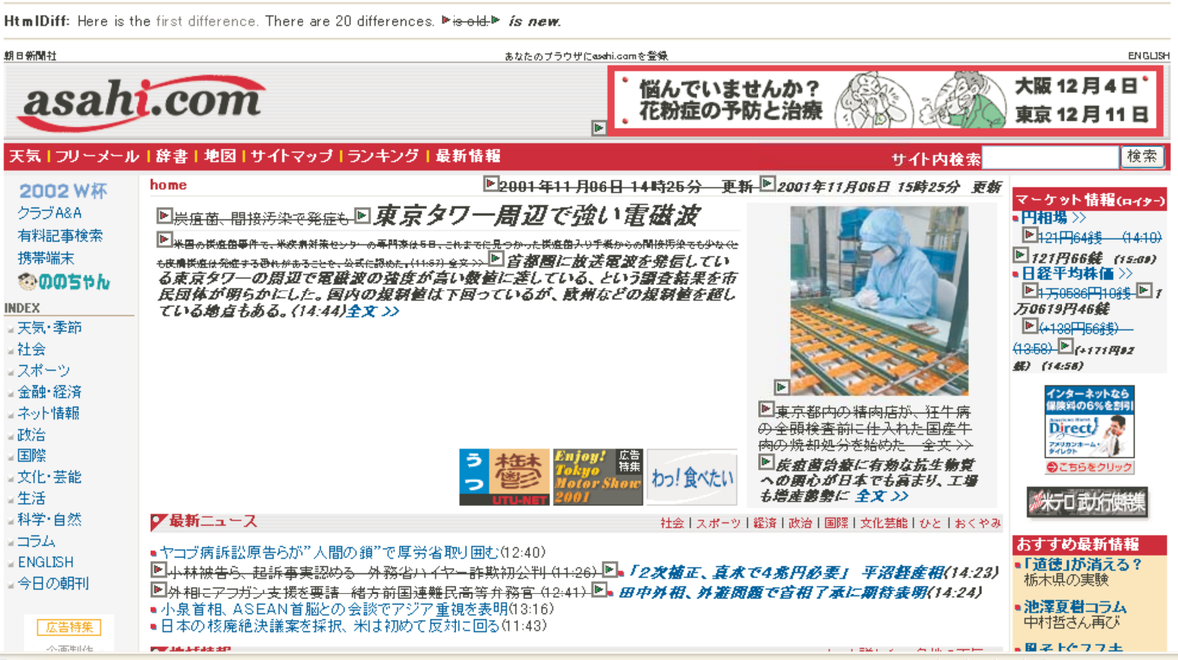
2.1 Monitoring Information Sources

An important feature of Web information sources is that they are frequently updated. The frequency of updates depends on the type of information source. For example, top pages of many university Web sites are updated about once a week. Those of news sites are updated more frequently at several times an hour. Moreover, Web pages that carry information about stock market, sports, auction, highway traffic are updated more frequently at once a few minutes. An active information gathering system is expected to gather information efficiently from such dynamically updating information sources.

AIDE (AT&T Internet Difference Engine) [3] is a tool, which has been developed at AT&T, to track changes of Web pages. As a component of the tool, HtmlDiff[1] has been developed as a publicly available software that compares two Web pages and displays

[1] http://www.research.att.com/ douglis/aide/

the differences. It highlights the difference by using deleted text for data struck out and italic font for data added as shown in Figure 2.



HtmlDiff: Here is the first difference. There are 20 differences. ~~is old.~~ *is new.*

朝日新聞社　あなたのブラウザにasahi.comを登録　ENGLISH

天気 | フリーメール | 辞書 | 地図 | サイトマップ | ランキング | 最新情報　サイト内検索　検索

2002 W杯
クラブA&A
有料記事検索
携帯端末
ののちゃん

INDEX
天気・季節
社会
スポーツ
金融・経済
ネット情報
政治
国際
文化・芸能
生活
科学・自然
コラム
ENGLISH
今日の朝刊

広告特集

home　~~2001年11月06日 14時25分　更新~~ *2001年11月06日 15時25分　更新*

~~炭疽菌、間接汚染で発症も~~ *東京タワー周辺で強い電磁波*

~~米国の炭疽菌事件で、米疾病対策センターの専門家は5日、これまでに見つかった炭疽菌入り手紙からの間接汚染でも少なくとも皮膚炭疽は発症する恐れがあることを、公式に認めた。(11:57) 全文 >>~~ *首都圏に放送電波を発信している東京タワーの周辺で電磁波の強度が高い数値に達している、という調査結果を市民団体が明らかにした。国内の規制値は下回っているが、欧州などの規制値を超している地点もある。(14:44)*全文 >>

~~東京都内の精肉店が、狂牛病の全頭検査前に仕入れた国産牛肉の焼却処分を始めた　全文 >>~~ *炭疽菌治療に有効な抗生物質への関心が日本でも高まり、工場も増産態勢に* 全文 >>

マーケット情報(ロイター)
円相場 >>
~~121円64銭　(14:10)~~ *121円66銭　(15:00)*
日経平均株価 >>
~~1万0586円10銭~~ *1万0619円46銭*
~~(+138円56銭)　(13:58)~~ *(+171円92銭)　(14:56)*

最新ニュース　社会 | スポーツ | 経済 | 政治 | 国際 | 文化芸能 | ひと | おくやみ

- ヤコブ病訴訟原告らが"人間の鎖"で厚労省取り囲む(12:40)
- ~~小林被告ら、起訴事実認める　外務省ハイヤー詐欺初公判 (11:26)~~ *「2次補正、真水で4兆円必要」　平沼経産相(14:23)*
- ~~外相にアフガン支援を要請　緒方前国連難民高等弁務官 (12:41)~~ *田中外相、外遊問題で首相了承に期待表明(14:24)*
- 小泉首相、ASEAN首脳との会談でアジア重視を表明(13:16)
- 日本の核廃絶決議案を採択、米は初めて反対に回る(11:43)

おすすめ最新情報
- 「道徳」が消える？ 栃木県の実験
- 池澤夏樹コラム 中村哲さん再び

Figure 2: An Output of HtmlDiff

TopBlend [4] is a revision of HtmlDiff and is implemented in Java, though it is not publicly available when we write this paper.

DataFoundry[2] [5, 6] is developed at Lawrence Livermore National Laboratory to maintain a data warehouse by detecting changes of information sources. In this project, the database schema of scientific data sources is represented as a graph that can be used to detect changes of the data and the schema itself. In scientific database systems, demands of database schema changes occur frequently, and to meet demands by manual operations costs much. This project aims at updating the schema automatically by tracking changes of information sources.

INRIA is also developing a data tracking system called Xyleme[3] [7] for maintaining XML-based data warehouses.

CONQUER [8] at Oregon Graduate Institute and NiagaraCQ [9] at University of Wisconsin are database approaches for monitoring information sources by formalizing the task as continuous queries.

2.2 Integrating Information

There are a number of Web information sources that deal with a similar topic. For example, there are a number of news sites on the Internet. The contents of each site is slightly different because of the difference of editors and/or news sources. The timing of updating contents is also different.

On the other hand, some readers may wish to read their favorite articles as soon as possible as they appear on the site and others may wish to read articles from a wide range of viewpoints even collecting articles takes time. The preference on reading

[2] http://www.llnl.gov/casc/datafoundry/index.html

[3] http://www.xyleme.com/

news articles depends on the reader. To provide a better service to each reader, the system needs to appropriately collect articles from multiple news sites considering the collecting time and the redundancy of articles.

Integrating information sources of different type adds more value to each of the information sources [10]. For example, there are a number of movie related sites on the Web. A movie site provides information about directors and actors, a theater site provides information about movies currently showing, and a critique site provides information about reviews on movie. By integrating information from these sites, a system can reply to a query such as "show me a movie directed by Steven Spielberg with three stars currently showing in Tokyo," which cannot be replied by information from any sole one of three sites.

To achieve an information integration task like above, we need to consider the quality and the cost of information gathering [11]. Generally speaking, there is a tradeoff between the quality of information and the cost of gathering the information. For example, if we approximate the quality of information by counting the number of information sources from which the information is gathered, obtaining information with high quality takes much time.

A planning mechanism would be required to make a good balance between the quality and the cost. To this end, an information gathering agent called BIG (resource-Bounded Information Gathering) [12] is developed at University of Massachusetts.

3 Intelligent Tickers: Toward An Active Information Gathering System

When we observe frequently updating information sources, we notice that the updates occur bit by bit. For example, the top page of some news site may be updated every ten minutes or so, but the amount of an update is only a few lines in the Web page, though the total amount of updates becomes quite large because such a small update occurs many times.

In this paper, we call such an object that carries a small piece of information "a ticker," and propose the Intelligent Ticker system that collects tickers from a number of Web information sources and integrates them to support the user's decision making or problem solving.

The Intelligent Ticker consists of an information gathering module and an information integration module as shown in Figure 3.

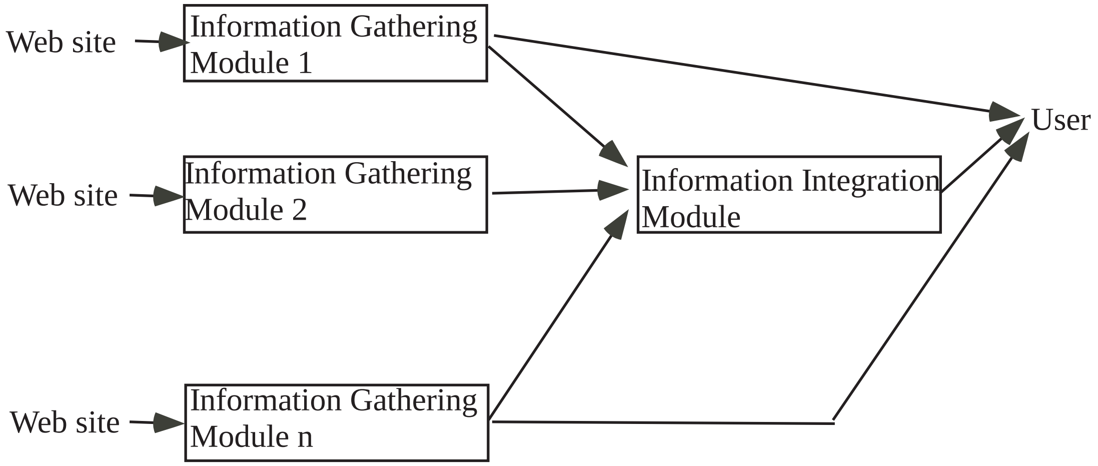

Web site
Information Gathering Module 1
Web site
Information Gathering Module 2
Web site
Information Gathering Module n
Information Integration Module
User

Figure 3: Overview of Intelligent Tickers

An information gathering module monitors a designated Web site and produces a ticker whenever an update occurs at the site. We assume a ticker is a part of Web page, and the user can show it directly on a Web browser.

The information integration module selects tickers sent from information gathering modules and integrates them to support the user's decision making and problem solving.

3.1 Information Gathering Module

Components of an information gathering module are shown in Figure 4. The Web access submodule fetches a Web page periodically from the designated Web site on the Internet and stores the sequence of the pages in the WebBase. The difference extraction module extracts the difference between two continuous pages in the sequence. This module uses the HtmlDiff mentioned in Section 2. The HtmlDiff shows differences by inserting tags into the original Web page, and the difference extraction module produces tickers from the output of HtmlDiff.

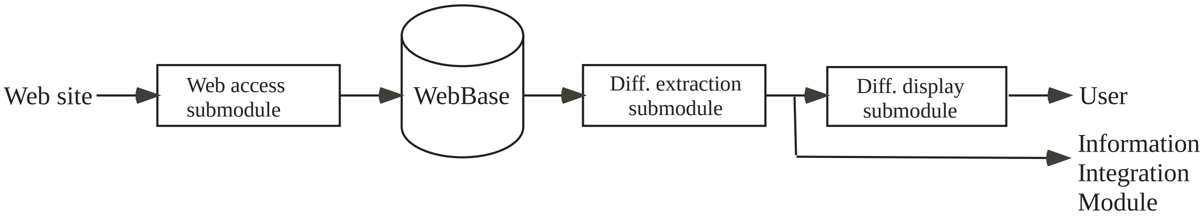


Figure 4: Information Gathering Module

The output of HtmlDiff can be analyzed as a tree structure as shown in Figure 5, for example, which is extracted from a news page shown in Figure 2. Each node represents a fragment of Web page consisting of text with the surrounding tags. Tags in a HTML document can be nested, so the document forms a tree structure. In this figure, a node depicted as a white thick circle represents a fragment which is actually updated, and should be left in the ticker. A node depicted as a white thin circle represents a fragment which is highly related to the updated fragment, such as the context or the category of updated article. We should leave such fragments also in the ticker to keep the updated fragment in the context. A node depicted as a gray circle is a fragment which should be left when the ticker is directly shown on a Web browser. Finally, a node depicted as a black circle is one which should be deleted.

How to build an analyzer that automatically categorizes the output of HtmlDiff into above 4 classes is an important and main research issue for developing the information gathering module.

A ticker, produced by an information gathering module, consists of the following elements.

- Object: An updated fragment of Web page. It is represented as text with surrounding tags.
- Time stamp: The time of update.
- Location: The URL of updated Web page.
- Context: The context of this ticker object.

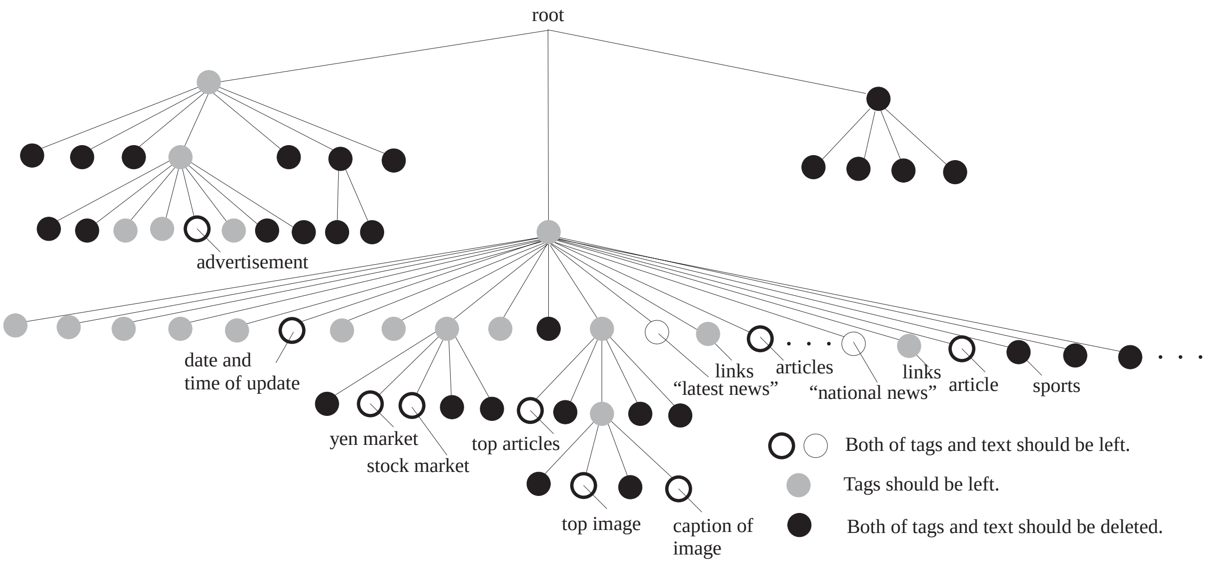


Figure 5: Analizing the output of HtmlDiff.

Each of information gathering modules needs to get access to the designated information source considering the frequency of updates. For example, it is useless to get access to an information source every hour when it is actually updated only once a day. Hence, we need to develop a Web access submodule with adaptability such that it learns the frequency of information source and dynamically change the interval of access.

3.2 Information Integration Module

The information integration module collects tickers produced by information gathering modules and integrates them to support the user's decision making or problem solving.

As the number of information gathering modules increases, the number of incoming tickers produced by the modules increases. Without selecting tickers, the information integration module may be overwhelmed by the incoming tickers, so we need to employ a ticker selection mechanism.

We use the ITT (Integration Template for Tickers), as shown in Figure 6, to integrate tickers. In this figure, choices of action that achieves a goal of traveling from Tokyo to Osaka are shown. We assume these choices are based on the user's preference. The user's first choice for traveling from Tokyo to Osaka are to take a bullet train or an airplane. The second choice is to take a night bus, and the last choice is to stay in Tokyo and to travel on the next day. By using this scheme, we know we need not collect tickers about the second and the last choices when the the first choice is satisfied.

For example, normally the information integration module collects tickers about bullet train and about airplane. We here assume that bullet trains are not available but airlines are available. Now let us assume that the module receives a ticker that airlines are not available, then the module begins to collect tickers about night bus. When it comes to know night buses are not available, the module begins to collect tickers about hotel and transportation of the next day. During such a process of information gathering, if it receives a ticker that bullet trains become available again, it stops collecting tickers about night bus and hotel. The information integration module dynamically changes the way of information gathering depending on the message of incoming tickers.

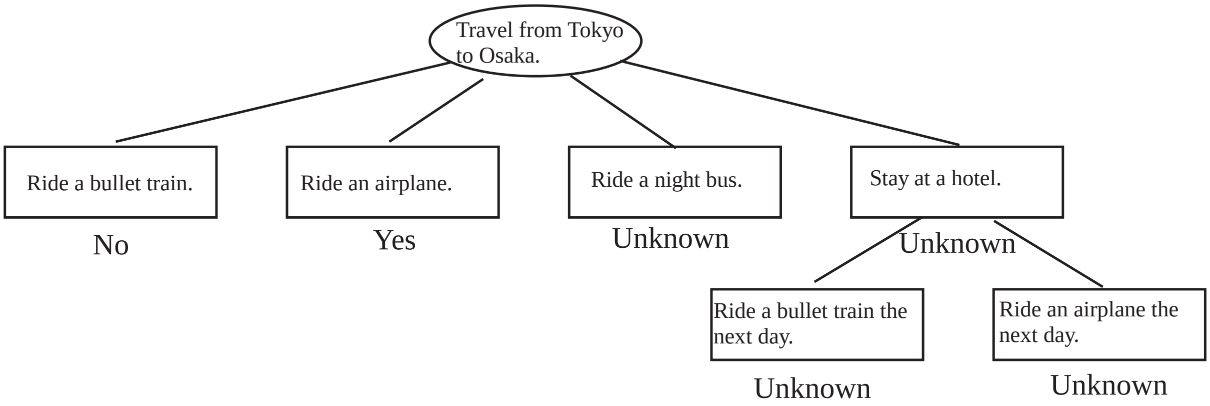


Figure 6: An example of Integration Template for Tickers.

When we consider an information gathering task, we can classify the information into the static one and the dynamic one. The dynamic information is frequently updated such as the availability of bullet train or airplane and is the main target of gathering. On the other hand, the static information is rather stable such as travel planning knowledge depicted in Figure 6 and is used to gather the dynamic information as mentioned above. Using the static information improves the performance of information gathering as discussed in [11].

Generally speaking, it is difficult to clearly define what is the static information and what is the dynamic information. The static information can be dynamic in a long run. For example, in the travel plan depicted in Figure 6, a night bus service may be abandoned or a cruise service between Tokyo and Osaka may start in the future. If so, the static information needs to be updated. Some static information may be updated directly by the user and others may be updated automatically by using collected information from the Web.

4 Conclusion

Active information gathering system gathers pieces of information from frequently updated information sources on the Internet and integrates them to assist the user in his/her decision making and problem solving task. It also works as an information gathering module of active mining system.

In this paper, we summarized required functionalities of active information gathering systems and proposed a new active information gathering system called Intelligent Tickers. At this moment, the system is still at the initial stage of concept development. We need to continue to develop the system as a useful system in the real world.

References

[1] M. Klein. XML, RDF, and Relatives. *IEEE Intelligent Systems*, 16(2):26–28, 2001.

[2] D. Fensel and M.A. Musen. The Semantic Web: A Brain for Humankind. *IEEE Intelligent Systems*, 16(2):24–25, 2001.

[3] F. Douglis, T. Ball, Y.-F. Chen and E. Koutsofios. The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web. *World Wide Web*, 1:27–44, 1998.

[4] Y.-F. Chen, F. Douglis, H. Huang and K.-P. Vo. TopBlend: An Efficient Implementation of HtmlDiff in Java. In *WebNet'00*. 2000.

[5] N. Adam, I. Adiwijaya, T. Critchlow and R. Musick. Detecting Data and Schema Changes in Scientific Documents. In *IEEE Advances in Digital Library*. 2000.

[6] T. Critchlow, K. Fidelis, M. Ganesh, R. Musick and T. Slezak. DataFoundry: Information Management for Scientific Data. *IEEE Trans Inf Technol Biomed*, 4(1):52–57, 2000.

[7] B. Nguyen, S. Abiteboul, G. Cobena and M. Preda. Monitoring XML Data on the Web. In *ACM SIGMOD*. 2001.

[8] L. Liu, C. Pu, W. Tang and W. Han. CONQUER: A Continual Query System for Update Monitoring in the WWW. *International Journal of Computer Systems, Science and Engineering*, 14(2):99-112. 1999.

[9] J. Chen, D.J. DeWitt, F. Tian and Y. Wang. Niagara CQ: A Scalable Continuous Query System for Internet Database. In *SIGMOD Conference*, pages 379–390, 2000.

[10] S. Yamada, T. Murata and Y. Kitamura. Intelligent Web Information System (in Japanese). *Journal of Japanese Society for Artificial Intelligence*, 16(4):495-502. 2001.

[11] Y. Kitamura, T. Noda and S. Tatsumi. A Dynamic Access Planning Method for Information Mediator. In *Cooperative Information Agents IV, Lecture Notes in Artificial Intelligence 1860*, pages 39–50, Springer, 2000.

[12] V. Lesser, B. Horling, F. Klassner, A. Raja, T. Wagner and S.X. Zhang. BIG: An agent for resource-bounded information gathering and decision making. *Artificial Intelligence*, 118(1-2):197-244. 2000.