http://ansinet.com/itj

ITJ

ISSN 1812-5638

# INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (4): 408-413, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.408.413


# Web Information Extraction Based on Visual Characteristics

Ruyue Tan
Room243 Apartment7, Harbin Institute of Technology,
Wenhuaxilu 2, Huancuiqu, Weihai, Shandong, People's Republic of China

**Abstract:** Due to the explosive development of Internet technology, the web is becoming the world's largest database of information, effective management and utilization of web information is currently a hot issue. This study mainly discusses the extraction of web information. Traditional web information extraction is mainly based on DOM tree and HTML tag analysis. Based on VIPS, the study proposes visual block positioning algorithm for webpage information extraction through induction webpage visual features and visual block feature information. It inputs the theme-based webpages and BBS webpages for VIPS, analyzes the output of VIPS and the VBT and then defines visual characteristics such as text density and link text density. The study puts forward a visual block positioning algorithm VBPA. It will position the theme information block to one VBT node, then extract the theme information. Experimental results show that the visual block positioning algorithm based on visual features is superior to the traditional web information extraction algorithm and has a higher quality of information extraction.

**Key words:** VIPS, visual pieces positioning, VBPA, Subject extraction, BBS information extraction

## INTRODUCTION

Rich in content exhibiting and interacting, the web conveys information visually (Cai *et al*., 2003a). Web visitors subconsciously partition the layout by visual characteristics to search more efficiently. Hence, in visual recognition, special information and visual features are significant. By analyzing visual information and layout features, the study manages to partition pages into blocks by VIPS (Cai *et al*., 2003b), effectively positions specific blocks and extracts further valuable information.

## RELATED WORKS

Information extraction methods can be classified into 2 classes: one is wrapper induction systems with delimiter-based rules; the other includes methods with syntactic/semantic constraints. There are also different taxonomies for extraction toolkits (Chang *et al*., 2006). For the first, Sarawagi classified wrappers into 3 kinds by tasks (Sarawagi, 2002): record-level, page-level and site-level wrappers. For the second, Laender categorizes it by wrapper-generating method (Laender *et al*., 2002): Languages for wrapper development (e.g., Minerva (Crescenzi and Mecca, 1998); HTML-aware tools (e.g., W4F (Saiiuguet and Azavant, 2001)); NLP-based tools (e.g., WHISK); Wrapper induction tools (e.g., WIEN); Modeling-based tools (e.g., NoDoSE (Saiiuguet and Azavant, 2001)). Vision-based extraction remains developing. Cai *et al*. (2003b) proposes a top-down independent label tree but fails to locate automatically in the study of Liu and Meng, (2006) extracts automatically but is too complex. By analyzing visual features of different areas and their general characters, VBPA is proposed.

## VIPS (VISION-BASED PAGE SEGMENTATION)

VIPS, proposed by Microsoft Research Asia, is an iterative top-down process. Its description will not be repeated again but is incorporated herewith by Cai *et al*. (2003b). VIPS parses a input page into a VBT (Liu and Meng, 2006), shown in Fig. 1. VBT has three characteristics: each node equals to

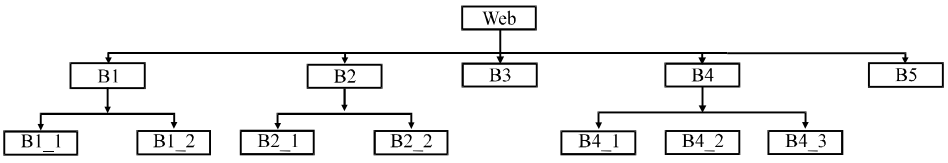


Fig. 1: An example webpage corresponding VBT

a visual block; each node equals to a visual area; VBT parent-child nodes are in containing relationship in corresponding page areas. Cai *et al*. (2003b) provides a segmentation algorithm but did not deal with extraction. Based on VIPS proposed by Cai *et al*. (2003b), the study proposes VBPA.

## VBPA AND WEB INFORMATION EXTRACTION

With visual blocks obtained by VIPS, information positioning and extraction can be done by VBPA.

**The eigenvalue of visual block B:** For each visual block B in the VBT, record its location, size, image information and text feature. Set the upper left corner vertex as the origin and the coordinates of the right bottom corner vertex as (Width, Height). Width and Height are the width and height of each block, the coordinates of center of each page block is (CenterX, CenterY). By VIPS, these can be got: distance from every B to the top margin of current page B_top, to the left B_top, position of the horizontal and vertical central axis L_land = B_top + 0.5 Height and L_protrait = B_left + 0.5 Width. Definitions of each B are:

**Definition 1:** The distance $\alpha$ between the horizontal central axis of B and the central axis of the parent node block of B is the absolute value of difference of the position of the horizontal central axis of B L_land and the position of the horizontal central axis of the parent node block of B L_fland:

$$\alpha = |L\_land\_L\_fland| \quad (1)$$

**Definition 2:** The distance $\beta$ between the vertical central axis of B and the central axis of the parent node block of B is the absolute value of difference of the position of the vertical central axis of B L_protrait and the position of the vertical central axis of the parent node block of B L_fprotrait:

$$\beta = |L\_protrait\_L\_fprotrait| \quad (2)$$

**Definition 3:** The semantic block B area S_B to the webpage area S_page is $\gamma$:

$$\gamma = \frac{S\_B}{S\_page}.$$

**Definition 4:** The pure text density $\lambda$ is the ratio of the pure text length in B to the area of B:

$$\lambda\_text = \frac{L\_textlength}{S\_B} \quad (3)$$

L_textlength is the length of the pure text in the visual block B.

**Definition 5:** The link-text density of B S_B is the ratio of the link-text length in B and the area of B:

$$\lambda\_link = \frac{L\_linklength}{S\_B} \quad (4)$$

L_linklength is the length of the link-text in the visual block B.

**VBPA for positioning and extraction theme content in theme-based website:** Theme-based pages extraction need to position and extract the theme block. Three specific steps are: generate a VBT from a theme-based page by VIPS; position theme block; extract theme information.

**Rules of visual block B:** In theme-based pages, theme area corresponds to a VBT node. Six rules of B are defined:

- **Rule 1:** The ratio of the link-text length of each direct child node of B L_linklength to the pure text length L_textlength satisfies:

$$\frac{L\_linklength}{L\_textlength} \geq 0.5$$

- **Rule 2:** The link-text density $\lambda$_link of each direct child node of B satisfies:

$$\log\frac{1}{\lambda\_link} \leq Z\_link$$

Z_link is the threshold determined by experiments on different sample webpages.

- **Rule 3:** The length-width ratio of each direct child node b of B is:

$$\varepsilon = \frac{b\_length}{b\_wide},$$

the width of B is B_wide, the height of B is B_high, then, $\varepsilon$<0.2 and:

$$\frac{\alpha}{B\_wide} > 0.6$$

or ε>5 and:

$$\frac{\beta}{B\_high} > 0.6$$

- **Rule 4:** The ratio of the distance β of the vertical central axis of B and the central axis of the parent node of B to the width of B'_wide the parent node block of B satisfies:

$$\frac{\beta}{B'\_wide} \le T\_disprotrait$$

T_disprotrait is the threshold determined by measurement

- **Rule 5:** The ratio γ of the area of B to the area of the webpage satisfies γ≥M_webregion, M_webregion is the threshold determined by experiments on different sample webpages

- **Rule 6:** If B is a theme block, its pure text density λ_text satisfies:

$$\log\frac{1}{\lambda\_text} \ge Z\_text$$

Z_text is the threshold determined by experiments on different sample webpages.

**Theme information block positioning algorithm:** The algorithm starts its layer-by-layer top-down iteration with the VBT root node; detailed as follow:

| Algorithm | | VBPA for Theme-based Webpage |
|---|---|---|
| I/O | | **Input:** VBT generated via VIPS |
| | | **Output:** VBT nodes which contains theme information |
| Steps | (1) | if the current node satisfies the rule 1, go to 4); |
| | (2) | if the current node do not satisfy the rule 1and satisfies the rule 2, go to 4); |
| | (3) | if the current node do not satisfy the rule 1and 2, but it satisfies the rule 3, go to4); |
| | (4) | label the child node satisfied the rules NOT theme information block; |
| | (5) | if the current node do not satisfy the rule 1, 2 and 3, iterate an unlabeled child of current node at next round; |
| | (6) | if current node do not satisfy the rule 4 and 5£¬turn to the next node of the current node, go to 1); |
| | (7) | if the current node satisfies the rule 4 and 5, then determine whether it satisfies the rule 6; |
| | (8) | if the current satisfies the rule 6, then label it as theme information block; |
| | (9) | return the node information of theme information block. |

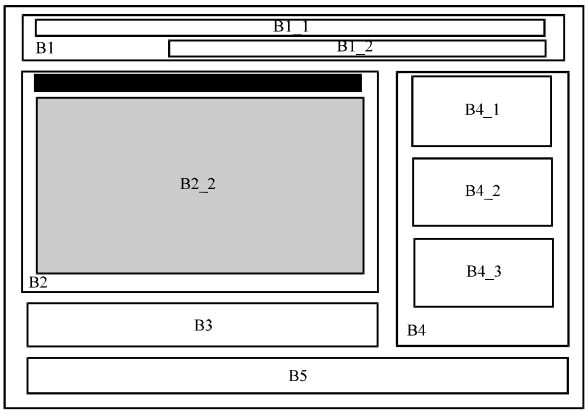


Fig. 2: The general structure of theme-based websites (LEFT)

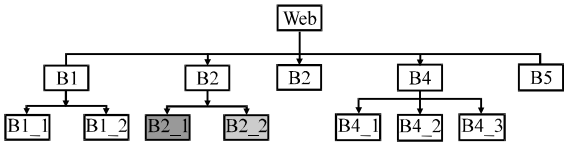


Fig. 3: Theme-based website VBT(RIGHT)

Figure 2 is to illustrate this process. B2, a parent block with title and body in the red frame, is positioned in the second round of VBPA. The next iteration positions theme block B2_2 which corresponds to a VBT node, shown in Fig. 3. Theme information can be extracted from the positioned block.

**VBPA for extracting posts speech information in the BBS website:** Based on the algorithm proposed in IV-B-2) the study provides VBPA to extract speech information from the positioned areas. Four specific steps are: (a) generate a VBT from a BBS page by VIPS; (b) position theme blocks; (c) cluster BBS speech areas in the theme blocks; (d) extract information from the set clustered in c). VBPA for BBS positions theme areas by VBPA for theme-based website and positions speech blocks by visual block feature clustering algorithm. The former VBPA, based on the latter, still needs further refined screening and positioning.

**Positioning the theme information block in BBS website:** Position theme blocks in the VBT generated by VIPS. Such blocks, possibly contains some speech areas, are both speech and users' information areas. Visual blocks which correspond to theme areas satisfy the rules defined in IV-B-1 but they have different threshold. BBS theme block positioning algorithm is given based on the

positioning algorithm proposed in IV-B-2). The VBT generated by VIPS is the input and the positioned VBT nodes is the output. See specific steps in IV-B-2.

**Visual block similarity of the posts in the BBS website theme area:** The theme block positioned in IV-C-1 has many child blocks with different content. Consider two child blocks B1 and B2 with areas S (B1), S (B2), respectively, areas of their parent blocks B1' and B2' are S (B1'), S (B2'), respectively, let $s = S(_B1) + S(B_2)$ and $s' = S(B_2') + S(B_2')$. Definition are as follow:

**Definition 6:** The position similarity $\delta$ of blocks B1 and B 2 equals to the ratio of the distance between B1, B2 and the left margin of its corresponding parent block, i.e.:

$$\delta = \frac{L\_lB1}{L\_lB2}$$

weight $\delta\_w = 0.3$

**Definition 7:** The area similarity $\eta$ of blocks B1 and B 2 satisfies:

$$\eta = \frac{s\_min}{s\_max};$$

its weight $\eta\_w$ is the ratio of the area sum s of B1 and B2 to the area sum s' of their parent $B_1'$ and $B_2'$, i.e.:

$$\eta\_w = \frac{s}{s'}$$

**Definition 8:** The number of content letters in B1 and B2 are C_num ($B_1$), C_num ($B_2$), respectively. Let C_min = min {C_num ($B_1$)}, C_max = max {C_num ($B_1$, C_num ($B_2$)}, content similarity $\theta$ satisfies:

$$\theta = \frac{C\_min}{C\_max}$$

the content weightis $\theta\_w$ the ratio of the text area sum s_t of B1 and B2 to the areas sum s of B1 and B2, i.e.:

$$\theta\_w = \frac{s\_t}{s};\ s\_t = S\_text(B_1) + S\text{-}text(B_2)$$

**Definition 9:** The similarity sim (B1, B2) of visual blocks B1 and B2 satisfies Eq. 5:

$$sim(B1, B20) = \delta \times \delta\_w + \eta \times \eta\_w + \theta \times \theta\_w \quad (5)$$

**Definition 10:** The similarity sim (Q, B) of visual blocks set $Q = \{Q_1, Q_2, \ldots Q_n\}$ and visual block B is the average value of the similarity of visual block B and every element in Q, i.e.:

$$\varepsilon = \sum_{i=1}^{n} sim(Q_i, B);$$

Q_max is the set who has the greatest similarity with B, it satisfies:

$$sim(Q\_max, B) = \frac{\varepsilon}{n}.$$

**Visual block clustering algorithm:** Process theme area blocks with level-by-level iteration clustering algorithm by similarity. This algorithm will aggregate all speech areas into a group in the first iteration round; in the following iterations, it will extract BBS page content information of different granularities shown as follow:

Algorithm: Visual block clustering algorithm in BBS website

I/O
- Input: VBT generated by VIPS
- Output: VBT nodes containing the BBS theme blocks

Steps
(1) Select a visual block from the first-level child nodes to build initial set Q which has only one element.
(2) FOR Select one visual block B which has not been marked as processed from the first-level child nodes
Calculate the similarity sim ($Q_i$, b) of visual block B and the formed set Qi
IF sim (Q_max, B)>M_sim (the threshold is obtained by experiments)
Put B into set Q_max
ELSE Built a new set which has only one element B
Mark B as processed, repeat the steps above until each first-level child node is processed.
(3) Select one child nodes from set Q; repeat (1), (2) until each one is processed. The algorithm is finished.

Theme block can be positioned by VBPA in IV-B-2), shown in Fig. 4. VBT node B2 is shown in Fig. 5, in gray

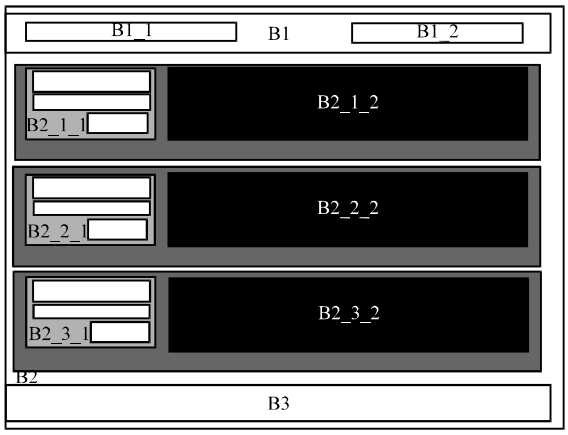


Fig. 4: BBS-type structure of webpage (LEFT)

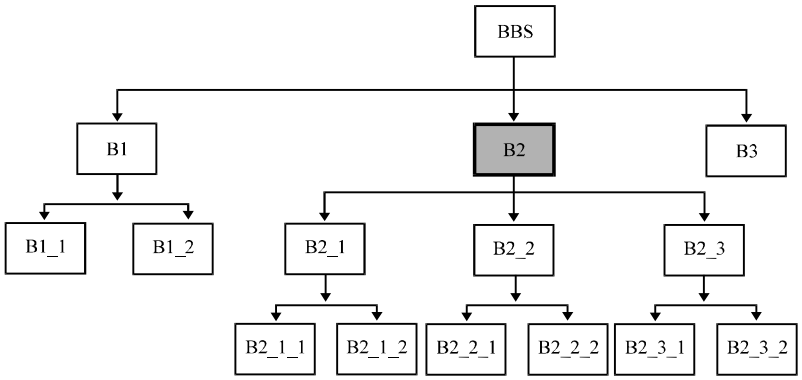


Fig. 5: BBS-based Web VBT(RIGHT)

Table 1: A sample of BBS-based Web information extraction results

| Content information set | | | | | |
|---|---|---|---|---|---|
| User name | Avatar | User data | Content | Time | … |
| Srelex | Ventris.jpg | Joined: 2010-01-20 09:33pm Posts: 1326 | Who else has heard of … | 2010-11-19 11:06am | … |
| Srelex | Ventris.jpg | Joined: 2010-01-20 09:33pm Posts: 1326 | Also, here's the first still.… | 2010-11-19 01:06pm | … |
| Manus Celer Dei | 3608.gif | Joined: 2005-01-01 07:30pm Posts: 1453 Location: I need you to relax your anus. | Huh. IMDB lists.… | 2010-11-19 02:49pm | … |
| … … | … | … | … | … | … |

frame. Cluster speech information similarities. Set B2_1, B2_2, B2_3 can be built in the first round iteration while set B2_1_1, B2_2_1, B2_3_1, B2_1_2, B2_2_2, B2_3_2 in the second and extraction of current granularity can be done. Specific position information in VBT is shown in Fig.5.

For one page http://bbs.stardestroyer.net/viewtopic.php?f=4&t=146025 of the BBS website http://bbs.stardestroyer.net/, the result of extraction after three rounds of iteration is shown in Table 1.

## EXPERIMENT AND RESULTS

**Hardware configuration:** Intel (R) Core (TM) Duo CPU P8700 at 2.53GHz×2, 2GB RAM, IDE: Visual Studio 2008. Extraction tests have been done on theme-based and BBS webpages, respectively.

**Experimental tests on theme-based webpage:** The data is collected from several famous Chinese portal-based sites and involves many domains like society, military, etc., shown in Table 2. Use label templates with HMTLParser to semi-automatically extract and mark the theme blocks in the 750 pages. The results are reported in three grades as excellent, good and poor. Excellent means all theme blocks are extracted precisely without interfere information; good means blocks are extracted but interfere information remains or information integrity is lower than 85%; others are poor. Table 3 shows the extracting results, where accuracy is (number of pages whose theme content is (precisely extracted/total number of pages)×100%.

Table 2: Experimental data set of theme-based pages

| Website | Theme type | No. of webpages |
|---|---|---|
| Sina | military, society, sports, finance, entertainment | 150 |
| Soho | military, sports, finance, public welfare, entertainment | 150 |
| NetEase | military, finance, science, sports, entertainment | 150 |
| Tencent | finance, science, sports, entertainment, fashion | 150 |
| Tom | sports, entertainment, society, automobile | 150 |

Table 3: Extraction results of theme-based pages

| | No. of input | Portion (%) | | | |
|---|---|---|---|---|---|
| Website | webpages | Excellent | Good | Poor | Accuracy (%) |
| Sina | 150 | 90.00 | 6.67 | 3.33 | 96.67 |
| Soho | 150 | 88.67 | 7.33 | 4.00 | 96.00 |
| NetEase | 150 | 83.33 | 14.00 | 2.67 | 97.33 |
| Tencent | 150 | 86.00 | 9.33 | 4.67 | 95.33 |
| Tom | 150 | 78.67 | 15.33 | 6.00 | 94.00 |
| Total | 750 | 85.334 | 10.532 | 4.134 | 95.866 |

This VBPA applies in supervisory systems of public sentiment in projects. The paper analyzes the blog-type results and accuracy in such systems. 200 pages from each site are analyzed, as in Fig. 6.

**Experimental tests on BBS webpage:** The test data is collected from several famous BBS forums and university forums in China, involving nearly 10 domains. It selects 5 types of BBS sites and pages of 4 theme kinds from each site and tests 540 pages chosen in total. Using label templates with HMTLParser to semi-automatically extract and mark speech blocks in the 540 pages and extracting by BBS VBPA. The results are shown in Table 4.

As is refined to speech block, it is not precise to experiment only on correctly positioned pages and

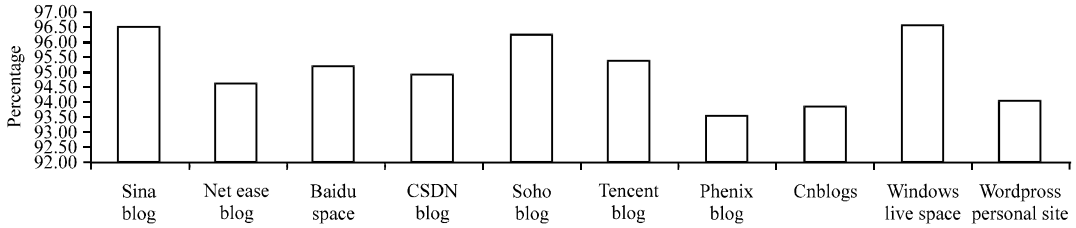


Fig. 6: Blog-type page extraction accuracies

Table 4: BBS-based page experimental results

| Website | No. of input pages | No. of correctly pages | Accuracy (%) |
|---|---|---|---|
| Discuz type | 80 | 77 | 96.25 |
| PHPWind type | 80 | 75 | 93.75 |
| DVBBS type | 80 | 73 | 91.25 |
| Forums as Sina, NetEase, Tencent, etc. | 200 | 189 | 94.50 |
| Forums as Tianya, Mop, etc. | 100 | 95 | 95.00 |
| Total | 540 | 509 | 94.259 |

Table 5: Comparison of two kinds of extraction algorithm

| Extraction basis | Total number of speech block in the input pages | No. of extracted speech blocks | No. of correctly extracted speech areas | Recall rate (%) | Accuracy (%) |
|---|---|---|---|---|---|
| VBPA | 1400 | 1287 | 1231 | 87.93 | 95.65 |
| HTML DOM tree | 1400 | 1182 | 1103 | 78.78 | 93.32 |

further calculation is demanded after speech-block-level extraction, where accuracy is the ration of correctly extracted speech areas to all extracted areas and recall rate is the ration of correctly extracted speech areas to all speech areas in input pages, both in percentage. The results are compared with traditional ones. As in Table 5, VBPA extraction has better recall rate and accuracy than that of HTML DOM tree extraction (9.15 and 2.33% greater, respectively), indicating the former has a higher recognition ratio and better performance.

## CONCLUSIONS

The study considers reading habits and visual features, proposes VBPA to extract information from pages of different types and structures, based on VIPS. It has better performance and higher accuracy and also proposes assumptions for large-scale algorithm in generic pages. Its complexity is lower than that of traditional ones while efficiency and accuracy are obviously enhanced. Further research will involve more page types, thus next step is to enhance accuracy, integrity and generality so that customized information of generic pages can be extracted by changing the thresholds and parameters.

## ACKNOWLEDGMENT

I would like to take this chance to express my sincere gratitude to my upperclassman, Xiao Yang, for his kindly assistance and valuable suggestions during my paper writing. My gratitude also extends to all the teachers who taught me for their kind encouragement and patient instructions. Special thanks to dear Sephiroth Chen, for the love, hope, courage and strength you have brought, bring and will bring to me. Last but not least, I would like to offer my particular thanks to my friends and family, for their encouragement and support for the completion of this study.

## REFERENCES

Cai, D., S. Yu, J.R. Wen and W.Y. Ma, 2003a. Extracting content structure for web pages based on visual representation. Proceedings of the 5th Asia Pacific Web Conference (APW' 2003), April 2003, Xi'an China, Springer, pp: 406-417.

Cai, D., S. Yu, J.R. Wen and W.Y. Ma, 2003b. VIPS: A vision-based page segmentation algorithm. Microsoft Technical Report, MSR-TR-203-79. http://research.microsoft.com/apps/pubs/default.aspx?id=70027.

Chang, C.H., M. Kayed, R. Girgis and K.F. Shaalan, 2006. A survey of web information extraction system. Inst. Electr. Electron. Eng. Trans. Knowledge Data Eng., 18: 1411-1428.

Crescenzi, V. and G. Mecca, 1998. Grammars have exceptions. Inform. Syst., 23: 539-565.

Laender, A.H.F., B.A. Ribeiro-Neto, A.S. da Silva and J.S. Teixeira, 2002. A brief survey of web data extraction tools. ACM SIGMOD Record, 31: 84-93.

Liu, W. and X. Meng, 2006. Vision-based web and data records extraction. Proceedings of the 9th SIGMOD International Workshop on Web and Databases, June 30, 2006, Chicago.

Saiiuguet, A. and F. Azavant, 2001. Building intelligent web applications using lightweight wrappers. Data Knowledge Eng., 36: 283-316.

Sarawagi, S., 2002. Automation in information extraction and integration. Proceedings of the 28th International Conference on Very Large Data Bases, August 20-23, 2002, Hong Kong, China.