A study on an efficient making method of detailed urban dataset by spatial integration of yellow-page data and digital maps for urban analysis

Yuki Akiyama
Graduate School of Frontier Science, The University of Tokyo
4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan
aki@iis.u-tpkyo.ac.jp

Ryoumei Onishi
Graduate School of Frontier Science, The University of Tokyo
4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan
aki@iis.u-tpkyo.ac.jp

Ryosuke Shibasaki
Center for Spatial Information Science, The University of Tokyo
4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan
aki@iis.u-tpkyo.ac.jp

Abstract: There were many researches of urban analysis with various ways in various disciplines. But information used often has only low spatial resolution especially when they use regional statistics, though the statistical data can over relatively large areas with homogeneous quality. For detailed analysis, some studies rely on field survey data that eventually have very fine spatial resolution, but they fail to cover an entire urban area or larger regions including interrelated urban areas.

The aim of this study is;

1) to develop a method of generating a urban dataset with fine spatial resolution that can cover an urban area or a whole national land territory itself, by spatially integrating detailed digital maps with yellow-page data and other company data,
2) to develop a new analysis method using the dataset generated with the method.

.

In this study, authors develop a method of highly accurate spatial integration of digital map data (Zenrin Co., Ltd Zmap-TOWNⅡ: digital map of Japan) with data that has attribute data (NTT Town page :telephone directory or yellow page of Japan). Z-map has polygon data describing building shapes. Town page has attribute data of resident's name or telephone number of residential houses and has tenant name or category of industries. The two datasets are spatially merged by address matching and name information identification. With this method, we can link not only Town page data but also various attribute data.

This urban analytical tool using the spatially integrated dataset targets to allow us to cover whole extent of Japan and multi-year with "detailedness" and flexibility. For example, we can track microscopic transitions of individual tenants near urban center to macro scale like distraction of industrial clusters with the same tool.

Keywords: urban analysis, GIS, utilization of existent information, address matching, yellow-page data, natural language processing, Perl programming

1. Background and aim of this study

"How to effectively research dynamic changes at urban space that has highest density of people and objects on earth" is the start of this study. We can acquire various information of urban space to collect many changes in cities and organize these information as spatial information, for example urban expansion, activation and declination of commercial avenue, tenant distribution etc. Recently about 50percents of world population are distributed urban areas, so it will become important theme all over the world that we get urban changes more closely by development new urban analytic datasets to unite existing information.

In this study, we focused on "Town page"(Yellow pages of Japan) to get urban changes. Town page has tenant information all over Japan. We can make new datasets that has detailed tenant information and accurate spatial information to unite Town page data and digital maps. And we can also research changes about each tenants to make multi-year dataset.

A biggest problem of this study is differences of character about each data. Development of new technology to unite difference spatial information is an aim of this study. Final aim of us is to unite difference spatial information almost 100percents accuracy. This study is a first step for this goal.

2. Properties of data

We use Town page (Yellow pages of Japan) to get tenant information and Zmap(digital map of Japan) to overlay Town page. Properties of data are as follows

1) About a Town page data

Town page is a yellow pages published by NTT (Nippon Telegraph and Telephone Corporation: the biggest telephone company in Japan). Tenants information are registered by self-return of tenant owners. Town page has these information.

1. Name of tenant
2. Address
3. Building name (Attention: There are same tenants that don't have this information)
4. Category assortment of businesses
   (Major category assortment number: 270 Minor category assortment number: 1120)

Town pages are popularly distributed as paper based, but we use digital data (csv.file) to unite digital maps.

2) About a Z-map data

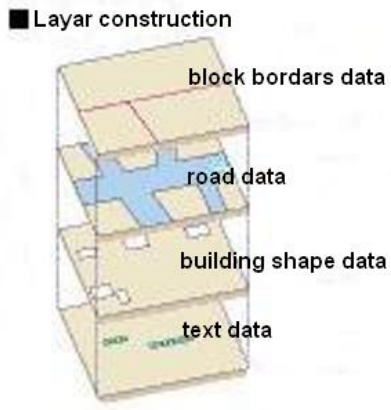


Figure2-1: Layer construction of Zmap (by ZENRIN HP)

Zmap(an official name is Zmap-TOWN Ⅱ) is a digital map marketed by ZENRIN CO., LTD. This map is a digital version of house map published by ZENRIN. A house map has block borders data, road shape data, building shape data and text data, and Zmap also has these data.

Zmap has layer structures like a figure2-1. There are layers of block borders data, road shape data, building shape data and text data attached on map. In this study, we unite Zmap and Town page data based on place information and object name information, so we use layers of building shape data and text data.

Also name information of Zmap are made by 280thousand of researcher to check door-to-door name plate of each house. Because of this way, some name errors happen between name data of Zmap and town page data that has faithful tenant names by self-return.

3. Method of object identification

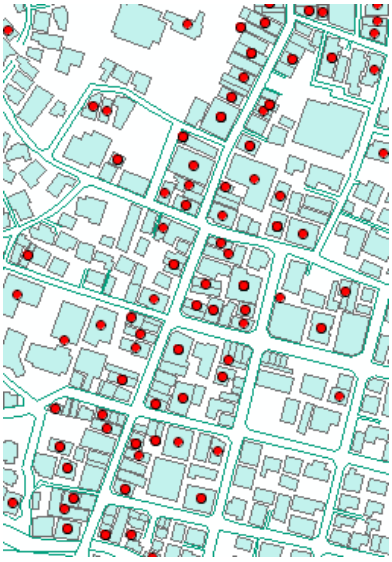


Figure3-1: Unity conclusion

This section explains how to unite information of Town page data and Zmap(the object identification). While we have explained that Town page and Zmap both have place information and object name information. So we judge passing status of object identification based on a size of unity about each place information and object name information.

Zmap is a digital map so it has digital data of place information. On the contrary, Town page has only text data of object's address. it doesn't have place information that can overlay on a Zmap. So we acquire place information to run an address matching based on Town page's text data of object's address. An address matching is a system that converts oblique place information like address and telephone number to direct place information like longitude and latitude. In this study we add information of longitude and latitude of each object to Town page data. By this operation, we can add object name information and place information to Zmap and Town page.

Then, we unite point data of Town page and nearest polygon data of Zmap building shape data by ArcGIS (figure3-1). We can unite Town page data and Zmap by this operation, but we can't identify each united objects as really same object. We only unite each Town page point and each nearest building

shape polygon. So we must make a method of examination that united objects are really same. We made this method based on name information and place information. Also we used "CSV address matching service"(http://www.tkl.iis.u-tokyo.ac.jp/~sagara/geocode/).

1) About place information

We defined average radius of buildings (equation1) based on building area to allow for various scale and shape buildings in Zmap and cases that point data of Town page slightly run out of building polygons' edge

We defined building polygon area of Zmap as S_z, average radius of buildings as R_z.

$$R_Z = \sqrt{S_z / \pi} \qquad (1)$$

Then, defined Euclidean distance between representative point of building polygon (centroid of building point) and point of Town page as de_{zt}, difference between average radius of buildings and Euclidean distance as equation2.

$$D_{zt} = \frac{de_{zt}}{R_z} = \frac{de_{zt}}{\sqrt{S_z / \pi}} \qquad (2)$$

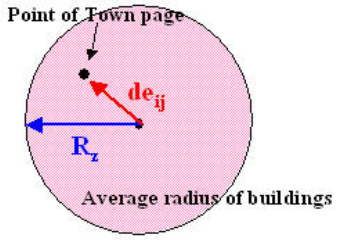


Figure3-2: Degree of position dissimilarity

This equation can evaluate that Euclidean distance between representative point of building and point of Town page (= de_{ij}) accounts for any percent of average radius of buildings (= R_z) (figure3-2). In this study, we defined this percentage as "Degree of position dissimilarity (DPD) ".

2) About object name information

We evaluate a similarity of object names to use the "n-gram" that is an evaluation index of similarity of linguistic structures. The n-gram is a statistics on frequency of appearance about N peaces of couples of logos or words that contiguously appear in texts. In this study, we made words blocks that are made from adjoining logos like figure3-3, and made percentage how many blocks coincide with all blocks. We defined this percentage as "Degree of name similarity ".

We defined $m_i^{(n)}$ and $n_{ij}^{(n)}$ as follows.

$m_i^{(n)}$: Number of blocks that has adjoining k logos extracted form text i.

$n_{ij}^{(n)}$:Number of $m_i^{(k)}$ that are equal to $m_j^{(k)}$.

Using these values, the n-gram that appear a degree of similarity about text i and text j is defined as equation3.

$$S_{ij}^{(n)} = \frac{n_{ij}^{(n)} + n_{ji}^{(n)}}{m_i^{(n)} + m_j^{(n)}} \qquad (3)$$

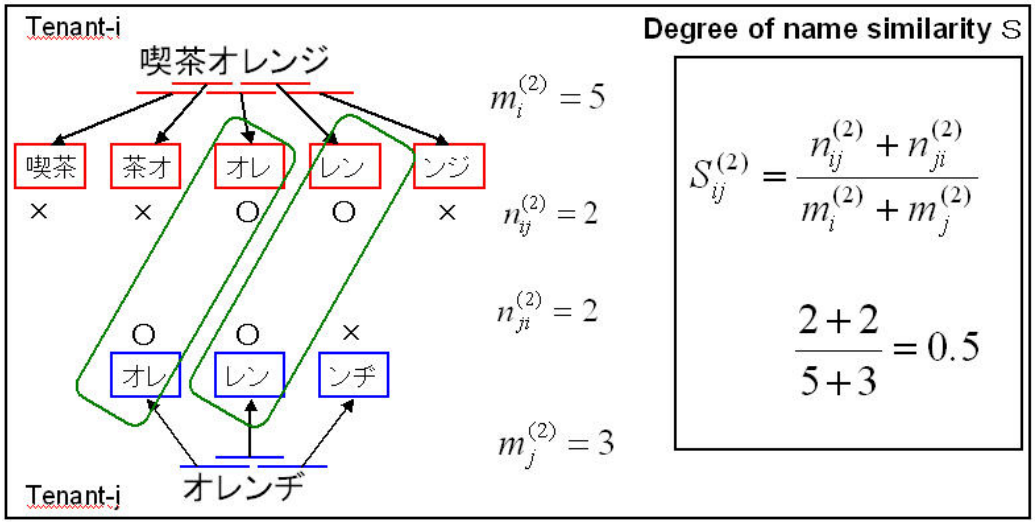


Figure3-3: Example of computation about two texts similarity (case of n=2)

In the case of n=1, It doesn't appropriate because some cases that matching logos appear happen even though each object were not same. So n=2 is a best value of n. But only this adjustment of value can't calculate an ideal degree of similarity in this study. Japanese has unique abbreviations, so the n-gram that n is 1 and the n-gram that n is 2 are classified the weight by 1:3, and we defined a degree of name similarity (= sn_{ij}) (equation4). We can check orders of logos by n-gram that n is 2 and can one-by-one check logos matches by n-gram that n is 1.

$$sn_{ij} = \frac{1}{4} S_{ij}^{(1)} + \frac{3}{4} S_{ij}^{(2)} \qquad (4)$$

If either text or both texts were one logo, degree of name similarity is defined by equation5.

$$sn_{ij} = S_{ij}^{(1)} \qquad (5)$$

And degree of name dissimilarity (DND) (=nsn_{ij}) is defined by equation6.

$$nsn_{ij} = 1 - sn_{ij} \qquad (6)$$

It has already introduced a case that Town page data have not only tenant names data, but building names data, and each data were collected by self-return of tenant owners. In contrast, name data of Zmap were collected by name plates. So there are many cases that one name data of Zmap doesn't match a tenant name of Town page but matches a building name data of it. So if object had tenant name data of Town page and building name data of it, we calculate each degree of name similarity and adopt higher one.

3) Method of synthesizing evaluation of each data

We made indices of position dissimilarity and name dissimilarity, but It's necessary to make united evaluation index to explain each indices. In this study we made two evaluation indices. One is a "Method of reference values comparing (MRVC)" and other is a "Method of principal component analysis (MPCA)". MRVC is a method that we set up reference values made by DND and DPD then compare DND and DPD with reference values, and only objects that can clear limits of reference values (DND and DPD below reference values can clear) pass. MPCA is a method

that we made degree of unity dissimilarity by the principal component analysis from each degree, then compare unity reference value of dissimilarity made from reference values by principal component analysis with degrees of each dissimilarity made from DND and DPD by principal component analysis, and only objects that can clear limits of unity reference value (degrees of each dissimilarity below unity reference value can clear) pass. Object synthesizing Accuracies of each method are pretty high. Accuracies are explained later (chapter5).

4. Development of object synthesizing method

We developed technologies to do object synthesizing automatically. In this study, we made an object synthesizing program by Perl, and use some free software. This system aims that anybody who want to use this system can use freely to use only free software and open source software.

1) Development of Perl program

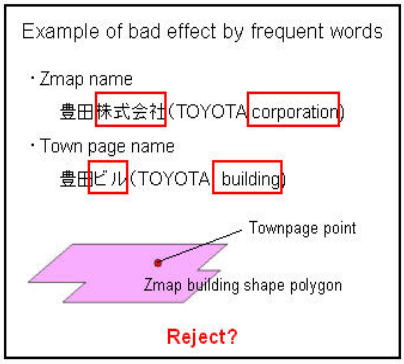
We developed programs by Perl that calculates DND and DPD, and check acceptances of object identification by MRVC and MPCA automatically. The Perl is a programming language made by Rally Wall, the most remarkable feature of this program is an excellent calculation function of texts. The n-gram is very complicated, and texts are also calculated very complicated. So we needed a programming language that is good at calculation of text. Because of this reason, we choose the Perl. Only 1000 lines of Perl program can calculate many complicate works, for example reading files, calculating DND,DPD and checking acceptances by MRVC and MPCA. And Perl is also a free software.

2) Utilization of free software

In this study, we used any free soft ware with Perl program. We introduce two free softwares in this section.

First free software is “RankWord”. A function of this software is an extraction of frequent words form texts with frequent ranking. This software divides letter strings into single words by morphological analysis. There are many cases that name information have words that don’t relate to proper names of tenant or building names (figure4-1). Because of this effect, there are cases that an object that should be checked as passing is checked as rejection (figure4-2). So before calculating by Perl program, delete frequent words.

Second free software is “Yomiyomi”. “Yomiyomi” means “Read read” in Japanese. A function of this software is a conversion Kanji and Katakana to Hiragana. Japanese uses three characters. There are Kanji, Katakana and Hiragana. Katakana and Hiragana are phonograms, and these phonograms express how to read Kanji. Japanese uses all three kinds of words in one text. So tenant names and building names are also expressed by these three kinds of words. Because of this reason, there are cases that one text that is converted Kanji to Hiragana coincide with the other text of Hiragana if one text is expressed by Kanji, and the other is expressed by Hiragana so one text doesn’t coincide with the other (figure4-3). This is a reason of utilization of Yomiyomi. Using this software, all Kanji convert to Hiragana, and accuracy of checking steps up more.



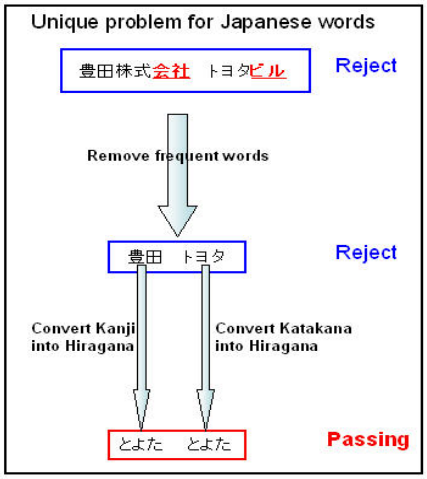


Figure4-1 (left): Example of frequent words
Figure4-2 (center): Bad effects of frequent words
Figure4-3 (right): Unique problem for Japanese words

5. Accuracy evaluation of object synthesizing method

In this section, we evaluate accuracies of object identification. A method of accuracy evaluation is that 1000 number of samples are selected at random, and we compared a manual check of object identification against samples with a check by Perl program. Also, errors of random sampling method are equation8 that are guided by equation7.

$$X = \frac{N}{\left(\frac{E}{k}\right)^2 \frac{N-1}{P(100-P)} + 1} \qquad (7) \qquad E = k\sqrt{\left(\frac{N}{X} - 1\right)\frac{P(100-P)}{N-1}} \qquad (8)$$

X: Number of samples N: Number of population E: Error margin
P: Result of assumption checks K: reliability coefficient (=1.96)

A accuracy comparing of manual check and Perl program tried at the Setagaya Word (Tokyo, Japan). Setagaya Word has about 33000 tenants. Based on a equation8, we can experiment with accuracy of 95.65percents in Setagaya Word. By the try of manual check, we can check objects with accuracy that are passing by manual check but are rejection by Perl program, or are rejection by manual check but are passing by Perl program. We evaluate accuracy of object identification with accuracy compared most accurate data by manual check with check results by Perl program.

Also we experimented accuracy evaluation not only about differing kind of data but also about same kind of data. We experimented to use Town page data in 2000 and 1995.

Table5-1: 4 groups of objects by MEVC division

○: Passing
×: Rejection

		Manual check		
		position	name	acceptance
Program check	Group1	×	×	×
	Group2	○	×	×
	Group3	×	○	×
	Group4	○	○	○

MEVC divides objects into 4 groups to evaluate accuracies (table5-1). 4 groups are like them. ① Degree of name dissimilarity (DND) is rejection and degree of place dissimilarity (DPD) is also rejection. ②Only DPD is passing. ③Only DND is passing. ④ DPD and DND are passing. Finally, group4 is judged as passing. A manual check judged all samples one-by-one by human. A Perl program check judged that DPD and DND clear reference values or no

Also best accuracy of reference values are that reference value of place is 40 percents and value of name is 70 percents.

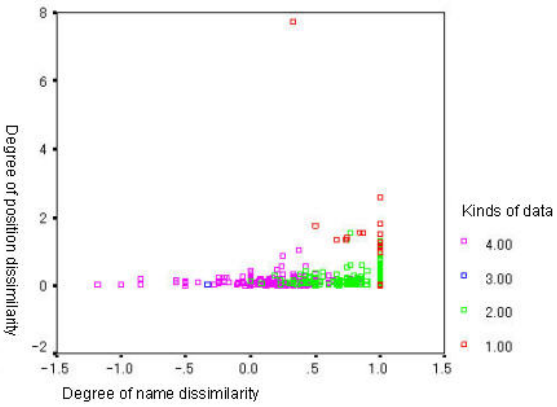
MPCA divides objects into 2 groups. By a principal component analysis, DPD and DND are combined to one new degree of dissimilarity, and we check that new degree clear a limit of a new degree of reference dissimilarity made from reference dissimilarity of name and place by a principal component analysis. 2 groups are like them. ①A new degree is passing. ②A new degree is rejection. Group1 is same as group4 of MRVC.

1) About a method of reference values comparing (MRVC)

MRVC is highly effective in object identification. MRVC accomplished 86.7 percents of accurate judgment about passing objects and 94.8 percents of complementation about manual checks (table5-1). But MRVC judgment failed around a reference value of DND (=0.7) where passing objects and rejection objects are mixed (fogure5-1 and 5-3). There are needs to be improved in these problems that MRVC can't catch passing objects over reference value, and judges rejection objects under reference value as passing.

2) About a method of principal component analysis (MPCA)

MPCA also has high accuracy of object identification. MPCA accomplished 86.2 percents of accurate judgment about passing objects and 95.2 percents of complementation about manual check (table 5-2), and accomplished 96.3 percents of accurate judgment about rejection objects, and accomplished 88.9 percents of complementation about manual check. We aimed to evaluate objects vaguely around reference value (=0.7) by principal component analysis, but principal component analysis isn't big effective in vague judgments, judging result of MPCA is similar with MRVC. Future prospects of MPCA are also how to deal with objects around reference value.



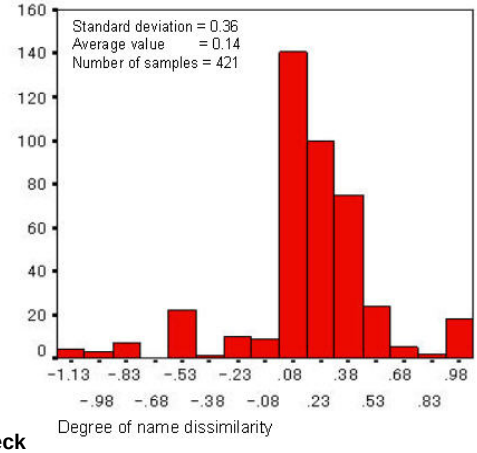
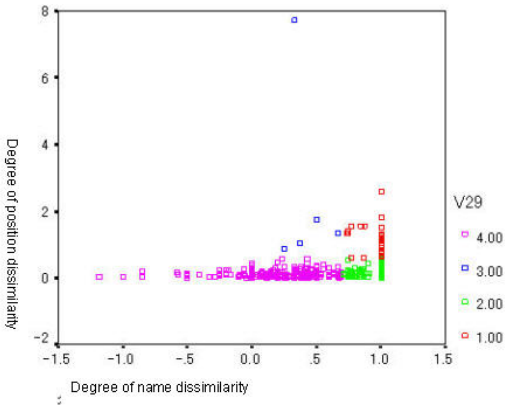


Figure5-1 (left): Distribution of DND and DPD by manual check
Figure5-2 (right): Distribution of DND about group4 by manual check



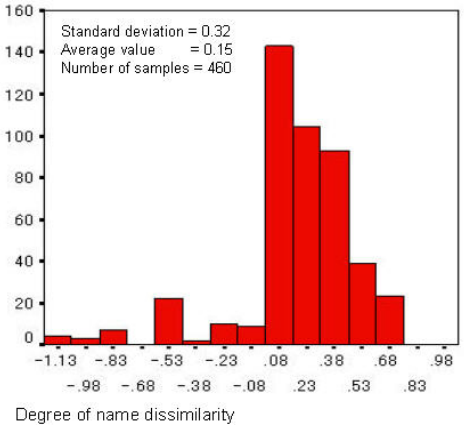
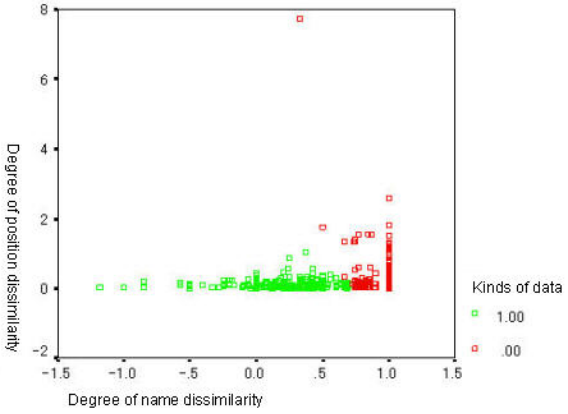


Figure5-3 (left): Distribution of DND and DPD by MRVC
Figure5-4 (right): Distribution of DND about group4 by MRVC



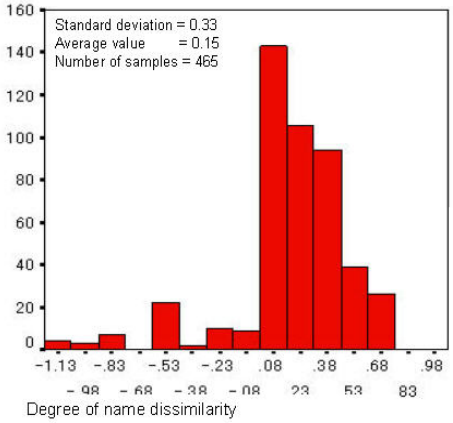


Figure5-5 (left): Distribution of DND and DPD by MPCA
Figure5-6 (right): Distribution of DND about group4 by MPCA

Table5-1: A list of accuracies about manual check and MRVC

			Manual check						
			Passing	Reject	Group4	Group3	Group2	Group1	
			421	579	421	1	557	21	ROC
Compareing method	Passing	460	399	61	399	1	60	0	0.867391
	Reject	540	22	518	22	0	497	21	0.959259
	Group4	460	399	61	399	1	60	0	0.867391
	Group3	5	2	3	2	0	0	3	0
	Group2	507	20	487	20	0	483	4	0.952663
	Group1	28	0	28	0	0	14	14	0.5
		ROC	0.947743	0.894646	0.947743	0	0.867145	0.666667	

Table5-2: A list of accuracies about manual check and MPCA

			Manual check		
			Passing	Reject	
			421	579	ROC
PCA	Passing	465	401	64	0.862366
	Reject	535	20	515	0.962617
		ROC	0.952494	0.889465	

How to read these tables?

Numbers out of heavy lines are numbers of objects extracted by each method. Numbers inside heavy lines are numbers of overlapping objects by each method. ROC means rate of complementarities. ROC is calculated as numbers of overlapping objects / numbers of objects extracted by each method.

3) Object identifications about same kinds of information

We achieved high accuracies of object identification by MRVC and MPCA even different kinds of information. Then we experimented object identification about same kinds of information. In this experiment, we used different years of Town page data. Years are 2000 and 1995. A method of experimentation is that we extracted 1000 samples and checked them by manual check and MRVC, then evaluated accuracies to compare a result of manual check with a result of MRVC. In this experiments, we checked only DND and rejected if DPD is more than 0 (Objects that DPD is more than 0 are only 43 in 1000 samples and 41 objects are rejected).

Table5-3: A list of accuracies about 2000' and 1995' NTT data

			Manual check		
			Passing	Reject	
			683	317	ROC
Compareing method	Passing	681	660	21	0.969163
	Reject	319	23	296	0.9279
		ROC	0.966325	0.933754	

We could accomplish highly accuracies than experiments about Zmap and Town page by MRVC. MRVC accomplished 96.9 percents of accurate judgment about passing objects and 96.6 percents of complementation about manual check (table 5-3), and accomplished 92.7 percents of accurate judgment about rejection objects, and accomplished 93.4 percents of complementation about manual check.

Using this system, we can make a tenants map of time-series changes like figure6-3.

6) Conclusion

Finally we summarize new knowledge and future prospects of this study.

1) New knowledge of this study

Previously, A technology of high accurate matching of different kinds of spatial information like Zmap and Town page had not build up. This study is a first step of technical development to prove this problem. New knowledge of this study are as follows

1. We developed program that can matches not only same kinds of existing spatial information but also different kinds of them with high accuracy.
2. This system can be constructed ArcGIS and some free software, so anyone who wants to use this system can promote an environment to use this system freely.
3. Anyone can match any information that is possible to be address matching (Information that has address or telephone number) to Zmap using this system.
4. Problems to match different kinds of spatial information were cleared.

2) Future prospects

Immediate aims are upgrading accuracies of matching by searching proper reference values of DND and DPD, and developing new index of dissimilarity etc, and development of Zmap and Town page united database all of Japan. In

the future, we will develop multi-yaer united database that can research time-oriented changes of objects. Immediate aims are as follows

① A developing of new reference values to evaluate DND and DPD comprehensively for better accuracy
② Grouping for conversion method English to Hiragana
③ Development an automatic system that operates whole software (Perl program, ArcGIS etc) in this system

Finally, we introduce same examples of spatial information made by this system.

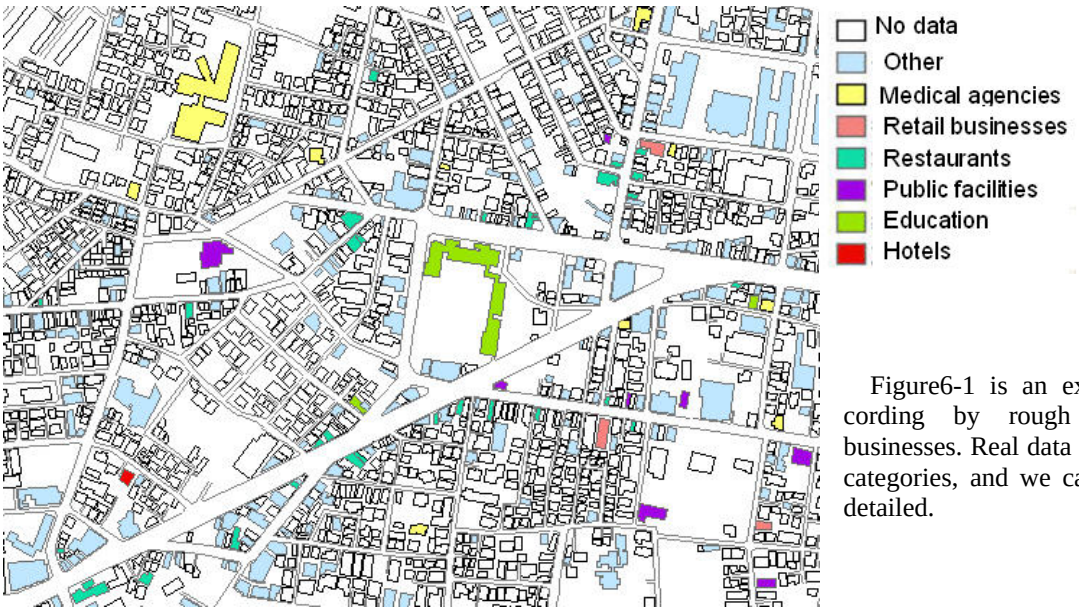


Figure6-1 is an example of color cording by rough categories of businesses. Real data has 200 kinds of categories, and we can analyze more detailed.

Figure6-1: An example of distribution of rough business categories

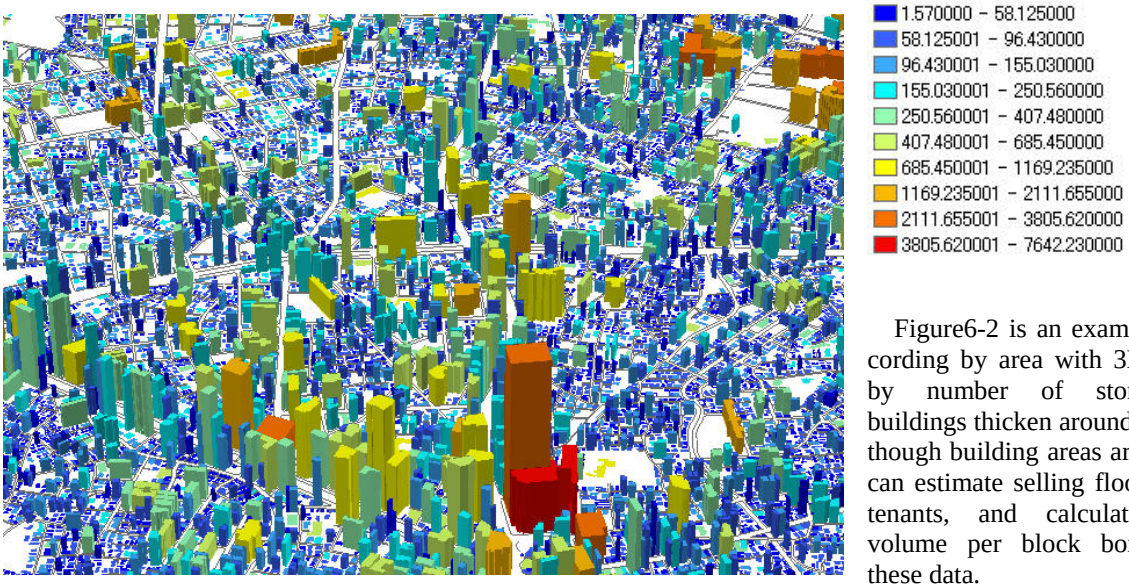


Figure6-2 is an example of color cording by area with 3D modeling by number of stories. High buildings thicken around main roads though building areas are small. We can estimate selling floor spaces of tenants, and calculate building volume per block borders using these data.

Figure6-3 is an example of color cording by time-series changes of Town page data (2000' and 1995'). Only 2000 are new appearance objects, only 1995 are extinct, change tenants changed tenants, and No change remain same tenants between 1995-2000. Using these data, we can see urban changes between macro scales (per cities or voluntary districts) and micro scales (per each tenants).

Figure6-3: An example of color cording by time series changes of Town page data

References

[1] Kaori Ito, 2001. Study of events in urban space, pp1-61
[2] Masayuki Asahara and Yuji Matsumoto, 2003. Japanese Named Entity Extraction with Redundant Morphological Analysis pp1-7
[3] Inoue Syoin, Architectural Institute of Japan, 1987. Examination and analytical method for architectural and urban planning
[4] Gizyutu Hyouronnsya, Syunji Mishima, 2004. CGI & Perl Pocket reference
[5] **URL**: CSV address matching service. Available at: http://www.tkl.iis.u-tokyo.ac.jp/~sagara/geocode/
[6] **URL**: Vector RankWord download site.
Available at: http://rd.vector.co.jp/soft/win95/util/se298263.html
[7] **URL**: Sarujie no siborikasu Yomiyomi download site
Available at: http://www.h2.dion.ne.jp/~georgia/