Method of extracting context from media data by using video sharing site

Satoshi Kondoh †, *Takeshi Ogawa* †

† NTT Network Service Systems Laboratories, NTT Corporation
Midori-cho 3-9-11, Musashino-shi, Tokyo 180-8585, Japan
E-mail: † {kondoh.satoshi,ogawa.takeshi}@lab.ntt.co.jp

ABSTRACT

Recently, a lot of research that applies data acquired from devices such as cameras and RFIDs to context aware services is being performed in the field on Life-Log and the sensor network. A variety of analytical techniques has been proposed to recognize various information from the raw data because video and audio data include a larger volume of information than other sensor data. However, manually watching a huge amount of media data again has been necessary to create supervised data for the update of a class or the addition of a new class because these techniques generally use supervised learning. Therefore, the problem was that applications were able to use only recognition function based on fixed supervised data in most cases. Then, we proposed a method of acquiring supervised data from a video sharing site where users give comments on any video scene because those sites are remarkably popular and, therefore, many comments are generated. In the first step of this method, words with a high utility value are extracted by filtering the comment about the video. Second, the set of feature data in the time series is calculated by applying functions, which extract various feature data, to media data. Finally, our learning system calculates the correlation coefficient by using the above-mentioned two kinds of data, and the correlation coefficient is stored in the DB of the system. Various other applications contain a recognition function that is used to generate collective intelligence based on Web comments, by applying this correlation coefficient to new media data. In addition, flexible recognition that adjusts to a new object becomes possible by regularly acquiring and learning both media data and comments from a video sharing site while reducing work by manual operation. As a result, recognition of not only the name of the seen object but also indirect information, e.g. the impression or the action toward the object, was enabled.

Keywords: media processing; context aware; video sharing site; web mining.

1. Introduction

Recently, a lot of research that applies data acquired from sensor devices to context aware services is being performed in the field on Life-Log and Sensor-Network. In various kinds of sensor devices, cameras and mikes are dealt frequently, because those sensors' data can include a lot of information. However, we should have converted these sensor's data to semantic data in order to use various general services, such as context aware services. Then, a variety of analytical techniques has been proposed to recognize various semantic data from sensors' raw data. In generally, these techniques use supervised learning which includes a training set that is generated by testees. Testees have to watch a huge amount of media data and have to input a huge amount of right semantic data manually. The works that generates a training set take a lot of time, and take a cost for collecting testees too. As a result, it is very difficult to change the training set, when it was made once. Therefore, the service can't reset the training data, even if semantic data that is mapped sensors' raw data will be changed as time is passed. The services such as recognition application often output wrong or old semantic data, because those services have to keep using inadequate training set.

Moreover, in the method of generating training set by collecting testees, mapped semantic data is often biased if testees are few. And, the variety of semantic data including not only content data but also related data is limited too, because the kinds of right semantic data are often limited.

In this study, we define semantic data as context data, and aim at a method that extracts context data from media data. As for the technical requirement of this purpose, the following four item are enumerated.

a) The time and cost for generating training set should be very low.
b) Training set should be changeable easily.
c) Training set should be generated by a lot of people.
d) Training set should include various semantic data such as not only content data but also related data.

Figure 1. Video sharing sites that allow users to give comments.

On the other hand, video sharing sites are remarkably being popular among web users. Among those sites, several new sites, i.e. "niconico-video"[1], "zoome", "spracia", etc. that enable users to give the media contents comments in detail has appeared, though conventional sites, i.e. "YouTube"[2], "ameba vision", etc. allowed users to watch the media contents in the past. Fig.1 shows the famous sites. "Niconico-video" enables users to give comments to the time position of the media contents freely. In this system, a media contents and comments can be regarded as a training set because the comments are mapped with the images in the media contents.

Then, we considered above video sharing sites as the means of generating training set. In those sites, the load of users isn't problem, because users attach comments to enjoy sharing their opinion about media contents. Therefore, requirement (a) will be satisfied by using media contents and comments in those sites. And, because these comments are frequently updated by users, requirements (b) will be satisfied by using those sites as the source of training set. Moreover, training set can reduce the influence of the bias by a specific user, because popular contents are watched and given comments by thousands of users. In other words, requirements (c) is satisfied by choosing the media contents. In those sites, users can freely attach various comments that include users' impression and information related to seen object in media contents, etc. Therefore, requirement (d) can be satisfied by using these various comments. However, it is difficult to use it as training set directly because variation of these comments is very numerous.

Then, we proposed the method that can extract appropriate words from comments in video sharing sites by filtering, and can map to the feature of image in media contents suitably. When the platform that has this function becomes possible, a retrieval service by using images as an input, and an archive service using video that users accumulated can be created easily, as shown in Fig.2. In this study, we confirmed the capability to make training set and to extract context data from this training set actually by implementing the prototype.

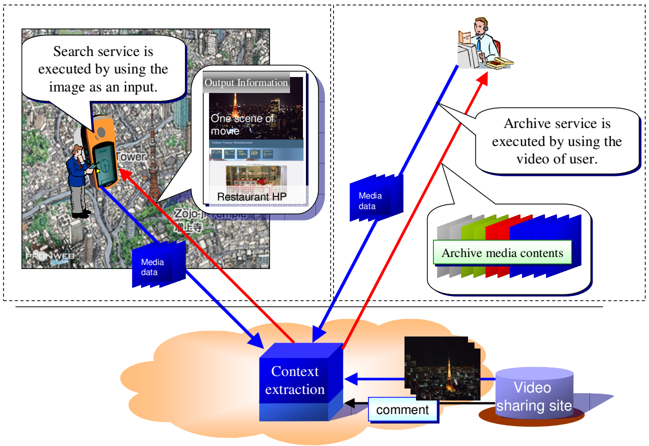


Figure 2. Service examples by Media Context Extraction.

2. Related works

In many study of image retrieval, the feature of media data, i.e. color histogram, spectrum, etc. is used. "Robust Media Search"[3] is a research for image retrieval by exact matching. This technique is robust towards changes of media contents by using feature that is not detail. According to this technology, application can retrieve images by a query of image. Context data can be acquired from the comments that are associated with the nearest image-frame by applying this technology if all beforehand the media contents are stored in DB. However, this method requires a huge storage and takes a sec-order time of search. Therefore, it is not suitable for extracting the context data from real-time media stream. And, if some similar media contents are exist in DB, comments that are associated with image-frame are dealt separately. Therefore output context data can't be merged automatically.

In the research to presume the camera position by the images, "Landmark database"[4] exists. This technique enables to estimate rough position by using Scale-Invariant feature. And, robust estimation of position can be achieved by voting above rough positions. In this technique, because we get only position data, the kinds of application which use this method are limited. For instance, image retrieval service required the other application in that the context data is mapped each position. Therefore, the cost for creating the map must be taken.

In the field of NW services, researches of media services are remarkable[5][6]. These researches aimed at the media processing on a network to improve the convergence of contents creation. However, because various processing includes recognition is still heavy for the server, light processing should be implemented to execute on the servers of NW nodes in the future.

3. Proposed technique

In order to solve the problem of the conventional research, we have proposed the system that enables extraction context data from video sharing sites. In this chapter, we explain an actual arrangement and functional details of this system.

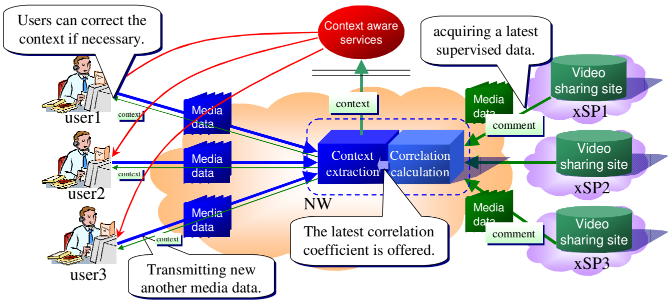


Figure 3. Media context extraction system.

3.1 Media context extraction system

The system is composed of the following two parts, which is "correlation calculation" and "context extraction", as shown in Fig 3. These functions are implemented in the server on a network. The server acquires the set that is composed of comments and media contents. At the same time, the server receives media data from users, and returns context data too.

3.1.1 Correlation calculation

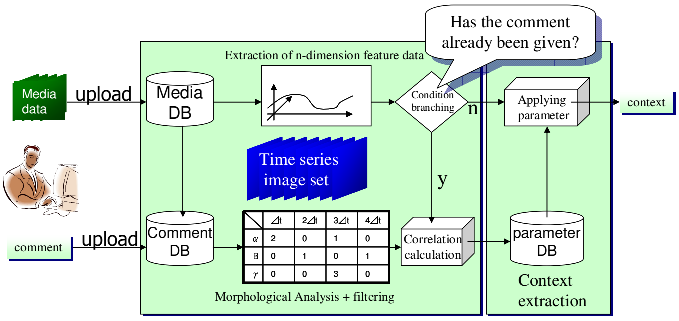
The correlation calculation function makes the training set for context extraction function. And it is composed of the following two processes.

Comment filtering - This processing executes morphological analysis to the sentences of the comments, and chooses the appropriate words. Then, the processing makes "word vectors" by collecting these words in specific time range. Finally, the processing gathers these vectors into a "word matrix".

Media Feature extraction - This processing extracts a feature data from all image-frames in media contents. After the processing makes "feature vectors" by same time range as the above-mentioned, it gathers these vectors into a "feature matrix". Next, it calculates a correlation coefficient between this matrix and above word matrix. Finally, it stores the correlation coefficient in "parameter DB". Fig.4 shows above-mentioned processing flow.

3.1.2 Context extraction

Context extraction function can apply above correlation coefficient to other media data. This function gets the correlation coefficient from parameter DB. At the same time, this function extracts feature data and makes a feature vector. Finally, it outputs semantic data as context data, by multiplying the correlation coefficient to the feature matrix.



	⊿t	2⊿t	3⊿t	4⊿t
α	2	0	1	0
B	0	1	0	1
γ	0	0	3	0

Figure 4. Processing flow in Media context extraction system.

3.2 Flow of processing in system

This paragraph explains the flow of use of the system. In this system, two flows exist. These two flows can work independently each other. One is following learning flow.

1. System accesses the video sharing site as well as the method seen by general users, and acquires the media contents and the comments.
2. System applies correlation calculation function to above data, and makes a correlation coefficient.
3. System stores above correlation coefficient in parameter DB together with the information that accompanies, i.e. filename, location data, etc.

Another is following context extraction flow. This is the flow that enables end-users to use this system. It is necessary to execute the learning flow beforehand several times to execute this flow.

1. Users' terminal transmits both media data and accompanied data to the system.
2. System gets accompanied information of received media data. And system chooses correlation coefficient associated with information that is the nearest to this information.
3. System applies above context extraction function to the media data, and gets context data.
4. System transmits the context data to the user terminal or application-servers that are registered beforehand.

3.3 Correlation calculation

3.3.1 Comment filtering

In the comments, there are profitable sentences for context data, and useless sentences exist. And, to improve generality as context data, system should output context data by the word than sentences. Then, system applies Morphological Analysis to the comment sentences to get the set of words, as shown in Fig.4. We use the library of Mecab[7] for this Morphological Analysis.

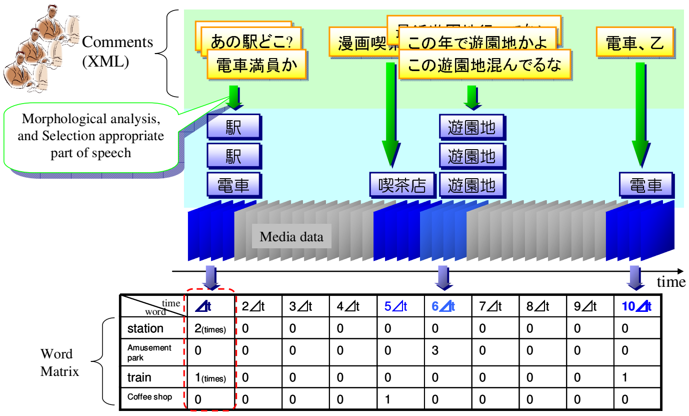
Next, system filters these words by a "part of speech". The passed "part of speech" is a noun, a verb, and an adjective. Moreover, system removes ASCII art and the word yell described by a specific character, i.e. "(TAT)", "ｷﾀ-", etc.

After system selected appropriate words roughly, it collects the words in the specific time interval that is determined beforehand, to make a "word vector". The word vector's row dimension shows the kind of above word. And elements show the frequency of the word generated in the interval of time. However, if system uses this frequency as a training set, a very excessive change will be learned. In other words, training set includes much noise. Actually, because the scene in the media contents is consecutive, the given comment influences at the time of neighborhood. Therefore, in our method, the model of generating the frequency by Gaussian distribution like the following expressions is used.

$$f(t_n) = \sum_w \sum_i F_{w\ \ i} \cdot e^{-\frac{(t_n - t_i)^2}{F_{w\ \ i} \cdot \sigma^2}} \qquad (1)$$

$F_{w\ \ i}$ is the original frequency of specific word w at time t_i, and σ is the fix standard deviation. On the other hand, the time interval has to be as small as possible because suitability by the learning becomes higher generally. However, generalization performance decreases when interval is set too small.

At the next step, system gathers these vectors into a "word matrix". In general, the word matrix is a sparse matrix. And this matrix still includes noise. Then, system executes dimension compression by PCA(Principal Component Analysis), and delete the data on the axis of low dimension. As a result, in the following processing such as media feature extraction can deal the small data. And the rest noise is reduced by compression. System stores basis conversion matrix that acquired at PCA, and brings the converted word matrix to media feature extraction processing.



time / word	⊿t	2⊿t	3⊿t	4⊿t	5⊿t	6⊿t	7⊿t	8⊿t	9⊿t	10⊿t
station	2(times)	0	0	0	0	0	0	0	0	0
Amusement park	0	0	0	0	0	3	0	0	0	0
train	1(times)	0	0	0	0	0	0	0	0	1
Coffee shop	0	0	0	0	1	0	0	0	0	0

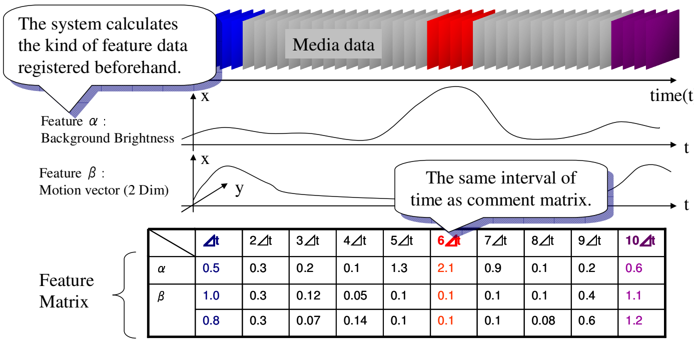
Figure 5. Processing of comment filtering.

3.3.2 Media feature extraction

In this process, system applies feature extraction functions that are registered beforehand to the media contents. The time interval is as same as above comment filtering processing. The sets of feature data that are extracted by different extraction functions are gathered into a feature vector, as shown in Fig.6.

In our method, though it is possible to adjust to all feature extraction functions, the kinds of function are limited for the accuracy of recognition actually. Concretely, edges and the corners are often used as feature data to be robust to the change in brightness. And, these feature data should be averaged timewise. Moreover, the change of feature data by the motion of camera angle should be canceled.

At the next step, system divides the feature vectors into the set of scene by detecting scene change. This change is detected by the change of pixels in the image. Then, system applies PCA to the set of feature vectors in the same scene. System calculates subspace information that is composed of the set of axes and range data, by using the eigen vectors and the eigen values.


The system calculates the kind of feature data registered beforehand.
Media data
time(t)
x
Feature α :
Background Brightness
t
x
y
Feature β :
Motion vector (2 Dim)
The same interval of time as comment matrix.
t
Feature Matrix

	⊿t	2⊿t	3⊿t	4⊿t	5⊿t	6⊿t	7⊿t	8⊿t	9⊿t	10⊿t
α	0.5	0.3	0.2	0.1	1.3	2.1	0.9	0.1	0.2	0.6
β	1.0	0.3	0.12	0.05	0.1	0.1	0.1	0.1	0.4	1.1
	0.8	0.3	0.07	0.14	0.1	0.1	0.1	0.08	0.6	1.2

Figure 5. Processing of Media feature extraction.

Next, system gathers the feature vectors into a "feature matrix". Then, system calculates correlation coefficient between above word matrix and the feature matrix, like the following expression.

$$\begin{bmatrix} W & b_0 \\ & \vdots \\ & b_{k-1} \\ \vec{0} & 1 \end{bmatrix} = \begin{bmatrix} & S & \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} & X & \\ 1 & \cdots & 1 \end{bmatrix}^{\dagger} \qquad (2))$$

Both W and $\vec{b} = (b_0 \cdots b_{k-1})^T$ are correlation coefficient. S is the word matrix, and X is feature matrix. In our approach, it is assumed that the relation between both is a linear model because this media context extraction system aimed at a platform for the various services. In general, system should specialize in some purposes to construct the nonlinear model that can improve recognition accuracy. However, it is no problem by selecting appropriate words in services that use the platform of this system, even if the output result includes some wrong words. And, even if result is seen a wrong word at a glance, it enables users to notice the information that users doesn't expect beforehand. Therefore, in this method, system uses linear model because the result can be output as wide as possible. On the other hand, because the load of processing by linear model is smaller than the nonlinear model, the linear model is suitable for the system for which the extraction function is used from a lot of services as a platform.

Then, system stores the correlation coefficient with above-mentioned basis convert matrix and accompanied data as a learning result, to the parameter DB. At the same time, system stores the information of subspace that determine applicable region.

3.4 Context extraction

This paragraph explains the method of context extraction by using above correlation coefficient. At first, when the system receives both media data and accompanied data, the system searches the appropriate learning result by using the accompanied data. The learning result that have the nearest accompanied data is selected, as shown in Fig.6.

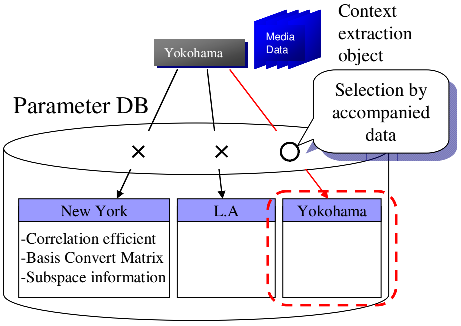


Figure 6. Selection appropriate the learning result on parameter DB.

At the same time, system makes a feature vector in the interval of time, by the method of above-mentioned media feature extraction processing. If this vector is in the subspace of the learning result, system multiplies correlation coefficient to the feature vector to get a word vector that is converted on the basis. At last, system multiplies the basis convert matrix to the word vector to get the word vector on original basis.

Finally, if values of the word vector are larger than the threshold that was determined beforehand, system output the corresponding words.

4. Experiments

In this method, it can be qualitatively confirmed to be able to reduce time because training set is automatically generated. However, the measurement of validity of output context data is the largest problem, because this method output various context data such as related information. Therefore, The judgment whether the output words are correct or wrong is very difficult. And, because a lot of words that show a similar meaning exist, the accuracy looks very low if comments given at the time which is different from the time of learning comments are used to compare with the output words.

Table 1: Specifications of experiment.

	Specification
Hardware	CPU: Core2Duo(2.0GHz), RAM: 4GB, OS: WinXP
Media Contents	Size: 512x384, fps: 30, Codec: VP6, Time: 120sec
Comments	The number of comments: 1000

Then, we examined generalization performance by cross validation, because we can compare in the limited set of words. For the experiment, we divided both the media contents and the comments equally at time, and used one block as input media data from users in that for the evaluation. The evaluation in the block is agreement between output words and original words simply. In this measurement, we used the experimental environment, as shown in table 1.

And, we used line and corner in an image as a feature vector. Fig.7 shows media contents which is used, and feature vector that is converted by the resized feature image. The size of feature image is 20x15 to prevent the dimension of feature vector from being huge.

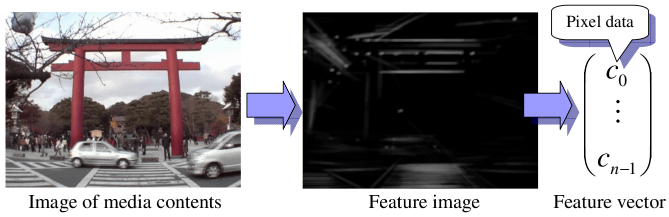


Figure 7. Media contents for the experiment and Feature vector.

In this experiment, we didn't use the feature data that correspond to the change, because the feature data that correspond to the change is not steady. It will be necessary to consider carefully about this feature in the future.

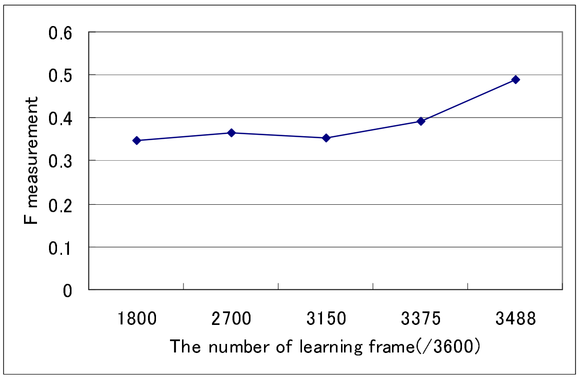


Figure 8. Generalization performance.

Fig.8 shows the accuracy by cross validation. Vertical axis is F measurement which is harmonic mean of precision and recall. Horizontal axis is the number of learning image-frame. In general, generalization performance is high if F measurement is high at the small number of image-frame. In the field of the retrieval, a system[8] whose F measurement is about 0.43 exists. In above figure, it can be confirmed to use it for the retrieval by collecting about 90% of the learning frames to be recognized. Therefore, the service in which the target images have been limited beforehand is suitable in this result. In the future, it is necessary to improve the accuracy at the few frames in order to enable other services that are used in unlimited scene to apply this technique too.

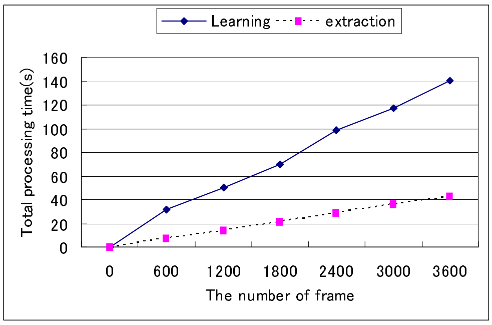


Figure 9. Time of learning and extraction.

A graph of the resultant processing times is shown in Fig. 9. In this graph, we measured two kinds of time, which are the learning time and the extraction time. Vertical axis is time of processing. And horizontal axis is the number of frame of media contents.

a) Learning time: The time taken for executing morphological analysis and filtering, extracting feature data, calculating correlation coefficient mainly.

b) Extraction time: The time taken for extracting feature data from users' media data, and multiplying the correlation coefficient to feature vector mainly.

As a result, we confirmed the extraction time is much smaller than the learning time. In general, because other method that is aimed at the media retrieval takes about sec-order for searching the query image from thousands of images, the proposal method can execute much quickly than those retrieval systems. And we confirmed that extraction processing spends about 100MB which is as small as memory resource used by a speech recognition server. In the future, because we expect to increase CPU performance and the expansion of the memory resource by the spread of 64bit OS, it is thought that those loads in a server become small very much.

On the other hand, we confirmed that the learning processing is low-speed, and takes a load which spends about 300MB on a memory. However, it is thought that there is no problem, because this processing has to do a little frequency, and is able to be executed by the background on the other divided server.

5. Conclusion

In this paper, we proposed a method of extraction context data from media data by training set that is generated by media contents and comments in video sharing site. And we confirmed the feasibility of the system. As a schedule for the future, we will experiment by more media contents and comments, and improve the mapping algorithm and the feature extraction function. In addition, we plan to create service that use this system, and verify effectiveness as the service of the output context.

6. REFERENCES

[1] Niconico video: http://www.nicovideo.jp/

[2] YouTube: http://www.youtube.com/

[3] S.Kimura,T.Kawanishi:"SPIRE:Similarity-based partial image retrieval guaranteeing same accuracy as exhaustive matching",MIRU2004, pp400-404,2004

[4] M.Susuki,T.Satoh:"Camera Position and Posture Estimation for a Still Image Based on a Landmark Database Using Scale-Invariant Feature",MIRU2007, pp660-665,2007.

[5] S.Takeuchi,Y.Kaneko,M.Yamamoto,M.Shibata:"A Program Production System using ID and File-data over IP Networks", SMPTE Motion Imaging Journal, vol.114,No. 3,pp.132-138,Apr.2005.

[6] S.Kondoh,T.Moriya,H.Ohnishi:"A study of integration of graph structures by filter-based WebAPI",IEICE society conference,B-19-21,2007.9.

[7] Mecab: http://mecab.sourceforge.net/

[8] Namazu: http://www.namazu.org/index.html.ja