Importance Estimation for Figures and Tables in Scientific Papers Based on Importance and Position of Referring Sentences

Takashi Hiraoka, Ryosuke Yamanishi, Yoko Nishihara, and Junichi Fukumoto

Abstract—**We propose a method of estimating the importance of figures and tables in scientific papers by propagating importance beyond media; from language to image. In scientific papers, language and image information coordinately enable the reader to easily understand the complicated contents in detail. The proposed method propagates this importance from the sentence to figures/tables in which the position of the sentence referring to a figure/table and surrounding sentences are used to evaluate the importance of figures/tables. We conducted an experiment on estimating the importance of figures/tables by assuming figures/tables used in a presentation poster are important. The experimental results indicated that the proposed method exhibited the highest mean of the average precision (MAP) compared to comparative methods focusing on the size and caption of the figures/tables. We believe that the proposed method is effective in supporting the creation of scientific presentation posters from papers.**

Index Terms—**importance propagation, application of natural language processing, creative support, poster design.**

I. Introduction

SCIENTIFIC papers are necessary for researchers to understand and share some knowledge and achievements with other researchers. However, most papers are highly technical, and it takes much time to read and understand the paper contents. The content of papers is efficiently presented in presentation posters and slides summarizing the important points. Presentation posters and slides of the paper are easy to understand for the readers but are hard to organize for the paper authors; knowledge and experience in design and a certain amount of times is necessary.

To support the presentation of scientific papers, several researches have proposed the methods of automatic generating presentation slides [1], [2], [3], [4]. Such support system would be helpful for most authors of scientific papers to prepare the presentation slides in oral sessions. More interactive sessions such as poster sessions have been recently increased at many conferences, therefore they would have a compelling need for a support system of generating "presentation posters." However, it is difficult to directly divert methods of automatic generating presentation slides to generating postres. Because the space of posters is limited, only necessary and sufficient sentences, figures, and tables must be selected to compose the attractive and intelligible scientific posters.

Qiang et al. proposed a method of automatically generating the presentation posters from the papers [5]. With their system, the important sentences and the layout are automatically determined using a machine learning mechanism, though the figures/tables used in the scientific poster are subjectively and interactively selected by the user. The figures/tables in a paper help the reader to easily understand the contents of the paper and are essential materials for a presentation poster. Only selected figures/tables, which are important in a paper, are included in the presentation poster. The figures/tables used in a poster are the key factors in evaluating the effectiveness of a poster. This may be one of the reasons that organizing posters is difficult. To organize posters, authors must consider the importance of the figures/tables, and have to choose a few appropriate ones several. Estimating the importance of figures/tables might be helpful in organizing presentation posters.

The ultimate goal of our research is fully automatic generation of presentation posters from scientific papers. As the elemental technology to achieve this goal, we propose a method of estimating the importance of the figures/tables propagating sentence importance.

II. Related Work

The purpose of this paper is multimedia summarization; a scientific paper is a type of multimedia contents consisting of language and image information. Previous studies have proposed video-summarization methods [6], [7], which are typically used for multimedia content. These methods to estimate the importance of image using dynamic image features are effective for variable images such as movies. However, the importance of tables in a scientific paper cannot be estimated using such methods since tables do not have image features but text and numbers. Accordingly, we developed a method of estimating both figures and tables, indirectly.

Estimation methods on the importance of the language information have been proposed, specially for document summarization [8], [9]. The importance of language information is well estimated in a certain performance. Most scientific papers mainly consist of sentences, and figures/tables are accessorily used to show some examples and details of the data. The proposed method is focused on such characteristics of scientific papers, and the importance of language information is propagated to figures/tables.

Manuscript received November 27, 2017; revised January 2, 2018.This work is supported in part by JSPS Grant-in-Aid for Challenging Exploratory Research #15K12103.

Hiraoka, T. is with the Graduate School of Information Science and Engineering, Ritsumeikan University, Nojihigashi 1-1-1, Kusatsu, Shiga 525-8577, Japan, E-mail: is0230ff@ed.ritsumei.ac.jp

Yamanishi, R., Nishihara, Y. and Fukumoto, J. are with the College of Information Science and Engineering, Ritsumeikan University, E-mail: {ryama, nisihara, fukumoto}@media.ritsumei.ac.jp

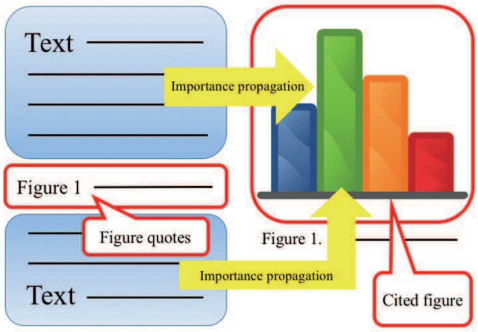

Fig. 1. Importance propagation from sentence to figure.

III. Proposed Method

As shown in Figure 1, the proposed method estimates the importance of a figure/table by propagating the sentence importance to the figure/table. Sentence importance is calculated based on the frequency of words in the paper and the similarity between each sentence. Sentence importance is propagated to the corresponding figure/table based on the positional information with the reference sentence for that figure/table while introducing the idea that the sentences related to the figures/tables should be positioned around that reference sentence. Thus, the sum of the importance for each reference sentence of a figure/table is assumed as the importance of the figure/table. The process of the method is detailed below.

A. Calculation of sentence importance

The importance of each sentence is calculated by using the word frequency. Previously, a sentence was parsed using a morphological analyzer. Then, only nouns are used in the importance calculation. We use the TextRank [10] is a graph-based ranking model to extract text information, which is used as the text extraction method in the existing study of generating presentation posters that Qiang et al. proposed. The importance of sentence i, $S(i)$, is calculated as the follows;

$$S(i) = \sum_{w \in \boldsymbol{W(i)}} n(w,i) \times f(w), \tag{1}$$

where, $n(w,i)$ and $f(w)$ denote the frequency of the noun w in sentence i and the entire paper, respectively. $\boldsymbol{W(i)}$ shows the set of nouns in the sentence i.

In scientific papers, logics and explanations are separately detailed in several sentences, and the importance of the sentence might be influenced by the importance of other sentences. So, the sentence importance is updated based on the similarity among sentences. The updating is capable of dealing with a stepwise logical explanation, e.g., a sentence details another explanation sentence. The similarity between sentences i and j, $R(i,j)$, is calculated using the cosine similarity. Then, the frequency of each word used in both sentences i and j is used as the vector value for each sentence.

The importance of sentence i, $Imp(i)$, is calculated as follows;

$$Imp(i) = S(i) + \sum_{j \in \boldsymbol{N}} R(i,j) \times S(j), \tag{2}$$

where N indicates the sentence set in the paper, and $R(i,j)$ shows the cosine similarity between sentences i and j. Accordingly, the sentence i has both the importance of itself and the importance of other sentences weighted by the similarity among sentences.

B. Propagation of sentence importance to figures/tables

The reference sentence itself is generally critically short, e.g. "*Fig. 1 shows the example of the proposed method*," and does not have specific important information. Truly important sentences related to the figures/tables (e.g., the explanation or detail of the figures/tables) should be before and after the reference sentences. The proposed method focused on the position of each sentence. The importances of the related sentences surrounding the reference sentence are propagated to that reference sentence, which is the substitue of the figure/table itself.

The reference sentence for each figure/table is retrieved from a paper. The design concept is that the closer to reference sentence of figure/table, the larger weight each sentence has. The weight for the i th sentence towards figure/table k based on the position of sentences i, $Pos_k(i)$, is calculated by the following equation, which can be regarded as a normal distribution with the average r and the standard deviation $= 1$;

$$Pos_k(i) = \sum_{r \in \boldsymbol{F_k}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(i-r)^2}{2}), \tag{3}$$

where, r and $\boldsymbol{F_k}$ denote the index of the figure/table and set of statements referring to the figure/table k, respectively[1]. The sentence importance $IMP(i)$ is propagated to the figure/table based on Equation (3) as follows;

$$PImp(k) = \sum_{i \in \boldsymbol{N}} Pos_k(i) \times Imp(i). \tag{4}$$

The range of $PImp(k)$ depends on the paper; thus, is normalized as follows;.

$$PImp'(k) = \frac{PImp(k)}{\sum_{k \in \boldsymbol{K}} PImp(k)}, \tag{5}$$

where, $\boldsymbol{K}$ denotes a set of figures/tables in a paper.

IV. Comparative method

In order to confirm the effectiveness of the proposed method on the importance estimation, we prepared two kinds of comparative method. Details of each comparison method will be described below.

[1] The parameters r and the standard deviation nevertheless should be calibrated.

A. Comparative method-Caption

Comparative method-Caption propagates the sentence importance without using the position information of figure/table quotes. Comparing comparative method-Caption and the proposed method, we would like to verify the validity of the location information that the figure/table is quoted when propagating the sentence importance. In the comparative method-Caption, the similarity between each sentence and the caption of each figure/table is considered instead of the point that figure/table is quoted and position information of each sentence when propagating the sentence importance to the figure/table. This model can be assumed as a model based on the idea "figure numbers and words frequently appear in the figure area," which is mentioned in the existing works [11], [12].

The sentence importance is calculated in the same way as the proposed method using equations (1) to (2). The similarity $CR(c_k, i)$ between the caption c_k of each figure/table and each sentence i in the paper is calculated by using the equation (6).

$$CR(c_k, i) = \frac{\displaystyle\sum_{w \in \boldsymbol{W(c_k)}, \boldsymbol{W(i)}} n(w, c_k) \times n(w, i)}{\sqrt{\displaystyle\sum_{w \in \boldsymbol{W(c_k)}} n(w, c_k)^2} \times \sqrt{\displaystyle\sum_{w \in \boldsymbol{W(i)}} n(w, i)^2}}, \tag{6}$$

In this equation, $n(w, c_k)$ indicates the appearance frequency of noun w appearing in the caption of the figure/table k. Also, $n(w, i)$ indicates the appearance frequency of noun w appearing in sentence i in the paper. $\boldsymbol{W(c_k)}$ indicates the noun set constituting the caption c_k. The importance degree $cCIMP(k)$ of the figure/table k in the comparative method-Caption is calculated as the equation (7).

$$cCIMP(k) = \sum_{i \in \boldsymbol{N}} CR(c_k, i) \times TIMP(i). \tag{7}$$

Similarly to the proposed method, we use the equation (8) to normalize the importance $cCIMP(k)$ of the figure/table k to the relative importance of the figure/table in the paper.

$$cCIMP'(k) = \frac{cCIMP(k)}{\displaystyle\sum_{k \in \boldsymbol{K}} cCIMP(k)}. \tag{8}$$

B. Comparative method-Size

Comparison between the comparative method-Size and the proposed method would let us confirm the effectiveness of propagating the sentence importance to the figure/table. Comparative method-Size does not propagate the sentence importance to the figure/table. In this method, the figure/table with the large area in the paper is assumed as the high important contents in the paper. The size of figure/table is used as the importance of the figure/table in the paper. The area $S(k)$ of the figure/table k is calculated by using the equation (9), assuming that the vertical length of the rectangle of the figure/table area in the PDF is $height$ and the horizontal length is $width$.

$$S(k) = height(k) \times width(k). \tag{9}$$

V. Experiment

To verify the effectiveness of the proposed method, we conducted an experiment on estimating the importance of figures/tables in scientific papers. We prepared two types of comparative methods: comparative method-Caption is based on the similarity between the caption for a figure/table and each sentence, and comparative method-Size is based on the size of figure/table.

We prepared 24 papers and their corresponding posters, i.e., paper-poster set, presented at the 30th Japanese Society for Artificial Intelligence[2] in the experiment. The figures/tables described in the posters were assumed as collectively important figures/tables for each paper. The proposed method and the two comparative methods were applied to estimate the importance of the figures/tables in each paper. The average precision was used as the evaluation index.

A. Experimental results

Table I shows the mean of the average precision (MAP) for the figure/table importance estimation with the proposed and comparative methods. The proposed method, comparative method-Caption, and comparative method-Size each exhibited almost 86, 79, and 79% MAP, respectively. The proposed method the most effective.

Table II shows beneficial relationship of the MAP among the methods. The proposed method was more effective than comparative method-Caption for 54% of the paper-poster sets and comparative method-Size for 42% of the paper-poster sets. This indicates the proposed method was more effective than the comparative methods for most of the paper-poster sets.

It seemed that the comparative method-Caption was effective for the papers in which the figures/tables had a certain sentence length. In the paper-poster sets used in the experiment, many papers had the short captions, e.g., "The proposed method" and "The results of the experiment." Since they were presented at a domestic annual conference without any review process, the caption was quite short. For such paper-poster sets, the effectiveness of comparative method-Caption would not work efficiently. Meanwhile, the proposed method focused on the relationships between the reference sentences for figures/tables and surrounding sentences. This characteristic is common for every scientific paper and does not depend on the content and writing style. The proposed method was therefore highly effective for most of the paper-poster sets.

The comparative method-Size did not take into account the relations between sentences and figures/tables. In scientific papers, sentences are commonly the main content and the figures/tables are used for additional materials to understand the detail of the content. Accordingly, the sentences and figures/tables have a certain relationship, which may be useful in estimating the importance of figures/tables; that is the point of the proposed method. From the results that the proposed method was more effective than comparative method-Size, it was suggested that using the relationship between sentences and figures/tables was more effective than using the size of figures/tables in estimating the importance of figures/tables.

[2] http://www.ai-gakkai/or.jp/jsai2016/

TABLE I
THE MEAN OF AVERAGE PRECISION OF FIGURE/TABLE IMPORTANCE ESTIMATION USING EACH METHOD CALCULATED FOR EACH PAPER (%)

Paper ID	Proposed method	Comparative method-Caption	Comparative method-Size
1	98.2	96.2	100.0
2	100.0	100.0	100.0
3	**100.0**	**87.7**	**87.7**
4	100.0	100.0	100.0
5	71.6	90.9	86.3
6	88.6	68.9	95.7
7	**64.5**	**34.0**	**47.4**
8	**70.0**	**63.9**	**47.8**
9	100.0	100.0	100.0
10	70.0	53.3	75.6
11	100.0	95.8	100.0
12	75.0	83.3	33.3
13	100.0	100.0	67.9
14	76.8	66.8	83.0
15	72.5	72.8	76.0
16	28.7	28.8	39.2
17	**100.0**	**41.7**	**50.0**
18	100.0	100.0	100.0
19	**79.6**	**78.6**	**76.0**
20	100.0	100.0	100.0
21	**98.2**	**90.9**	**75.5**
22	**79.6**	**66.0**	**59.6**
23	**90.9**	**82.6**	**87.4**
24	100.0	100.0	100.0
Average	86.0	79.3	78.7

TABLE II
BENEFICIAL RELATIONSHIP OF THE AP AMONG METHODS (%).

Relation	Comparative method-Caption	Comparative method-Size
Case of “proposed method $>$ comparative method”	**54%**	**42%**
Case of “proposed method $=$ comparative method”	29%	29%
Case of “proposed method $<$ comparative method”	17%	29%

TABLE III
THE RESULTS OF THE IMPORTANCE ESTIMATION AND THE FREQUENCY OF CITATIONS FOR EACH FIGURE/TABLE IN PAPER ID17 [13].

Caption number	Estimated importance	Citation frequency
Figure 2	**49 % (10393)**	**4**
Figure 1	**27 % (5694)**	**3**
Figure 3	12 % (2468)	1
Figure 4	8 % (1608)	1
Figure 5	4 % (915)	1

TABLE IV
THE RESULTS OF THE IMPORTANCE ESTIMATION AND THE FREQUENCY OF CITATIONS FOR EACH FIGURE/TABLE IN PAPER ID17 [14].

Caption number	Estimated importance	Citations frequency
Figure 9	15 % (10022)	4
Figure 10	12 % (7911)	3
Figure 5	10 % (6945)	2
Figure 8	10 % (6856)	2
Figure 1	8 % (5440)	2
Figure 11	**8 % (5344)**	**2**
Figure 12	**8 % (5249)**	**2**
Figure 7	7 % (5162)	2
Figure 3	**7 % (5138)**	**2**
Figure 6	6 % (4224)	1
Figure 2	**5 % (3112)**	**1**
Figure 4	4 % (2645)	1

B. Discussions with certain cases

We focused on certain cases and will discuss the experiment results in detail. The proposed method with the paper ID17 [13] showed a high MAP better than either comparison methods. On the other hand, the paper ID16 [14] showed a low effectiveness with the proposed method. We take these results up in the following discussions. Table III and IV each shows the result of the importance estimation by the proposed method and the frequency of the citation for each figure/table in the paper ID17 and ID16, respectively. Caption number in boldface represents the correct figure/table in this experiment, which were used in their presentation poster. The value in parentheses indicates the estimated importance $PImp$ before the normalization. The figures/tables are sorted in the order of the estimated importance.

For the figure/table cited several times in the paper, the proposed method calculated the importance by using the position information with each reference of the figure/table and summed the importance. Since, the figure/table with many citations tended to be estimated as the relatively high important contents. In the paper ID17 which result is shown in Table III, the figure with many citations in the paper was used in their poster. Then, the proposed method seemed to effectively work for the paper in which the importance of figure/table is highly related with the frequency of the citations. On the other hand, the paper in Table IV does not necessarily have high importance even for figures with many citations in the paper. It is considered that the proposed method is not effective for papers in which the frequency of citation does not influence the importance.

We focus on $PImp$ in each paper. The paper ID17 has $PImp$ gradually increased from figure 5 which is estimated as the highest important to figure 1 which is estimated as the lowest important On the other hand, in the paper ID16, figure 9 has a higher importance than the other figures but figure 10 to figure 3 in table IV showed substantially flat values. In the papers showing many figures as the examples of interfaces and processed images, there is no huge difference in the importance for each figure. The papers using figures for supporting to understand the contents seems to have different importance for each figure/table. That is, it is considered that the role of figure/table in the paper has a relation with the importance of the figure/table.

VI. Conclusion and Future Work

We proposed a method of estimating the importance of figures/tables in scientific papers. The proposed method estimates the importance of the figures/tables by propagating sentence importance. The sentence importance is calculated based on the word frequency and sentence similarity. The calculated sentence importance is propagated to the figures/tables while weighting the importance with the position relation between the reference sentence for the figure/table and surrounding sentences. Through an experiment on estimating the importance of figures/tables using 24 paper-poster sets, the proposed method exhibited higher MAPs than the comparative methods that are focused on the caption and size of figures/tables.

We believe that the importance of figures/tables can be applied to determine the size of figures/tables in presentation posters. The current method [5] uses the selection of the important sentences and the layout on the poster. By using both their method and our proposed method, a presentation poster can be automatically generated including figures/tables without any human hands. This combination of the two methods is for future work. Also, we will apply the idea of the proposed method for formulas which was not focused in this paper.

Acknowledgments

Asiss. Prof. Mitsuo Yoshida of the Toyohashi University of Technology gave advices to this work. We show our appreciation to the authors of the research presentation paper and the information presented at the poster-session of JSAI 2016, which are used in the experiment and analysis in this paper.

References

[1] K. G. Prasad, H. Mathivanan, and T. Geetha, “Document summarization and information extraction for generation of presentation slides,” in *Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing 2009*, 2009, pp. 126–128.

[2] R. Spicer, Y.-R. Lin, A. Kelliher, and H. Sundaram, “Nextslideplease: Authoring and delivering agile multimedia presentations,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 8, no. 4, pp. 43–48, 2012.

[3] P. Ganguly and P. M. Joshi, “Ipptgen-intelligent ppt generator,” in *Proceedings of International Conference on Computing, Analytics and Security Trends 2016*, 2016, pp. 96–99.

[4] D. Edge, J. Savage, and K. Yatani, “Hyperslides: dynamic presentation prototyping,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 671–680.

[5] Y. Qiang, Y. Fu, Y. Guo, Z.-H. Zhou, and L. Sigal, “Learning to generate posters of scientific papers,” in *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI’16)*, 2016, pp. 51–57.

[6] K. Miura, R. Hamada, I. Ide, S. Sakai, and H. Tanaka, “Motion based automatic abstraction of cooking videos,” *MIRU2002*, vol. 2, pp. 203–208, 2002,（三浦宏一，浜田玲子，井手一郎，坂井修一，田中英彦，“動きに基づく料理映像の自動要約手法” 画像の認識・理解シンポジウム（MIRU2002）論文集.）．

[7] T. Yao, T. Mei, and Y. Rui, “Highlight detection with pairwise deep ranking for first-person video summarization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 982–990.

[8] K. Zechner, “Fast generation of abstracts from general domain text corpora by extracting relevant sentences,” in *Proceedings of the 16th conference on Computational linguistics*, vol. 2, 1996, pp. 986–989.

[9] K. Hong and A. Nenkova, “Improving the estimation of word importance for news multi-document summarization,” in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 712–721.

[10] R. Mihalcea and P. Tarau, “Textrank: Bringing order into texts,” in *Proceedings of Conference on Empirical Methods on Natural Language Processing 2004*, 2004, pp. 404–411.

[11] J. Ichino, K. Mimaki, K. Yamaguchi, S. Kaki, I. Azuma, and S. Furuta, “Experiment in automatic extraction of chart information,” *IPSJ SIG Technical Reports*, no. 28, pp. 143–150, 2002,（市野順子，箕牧数成，山口和泰，垣智，東郁雄，古田重信，“図表検索のための図表情報自動抽出の試み” 情報処理学会研究報告.）．

[12] R. Takeshima and T. Watanabe, “Extraction of figures/tables-specific explanation sentences based on interdependencies between sentences and words,” *The Institute of Electronics, Information and Communication Engineers*, vol. 110, no. 85, pp. 43–48, 2010,（竹島亮，渡邉豊英，“文と単語の相互依存性に注目した図表説明文の抽出” 電子情報通信学会技術研究報告.）．

[13] S. Kumatani, T. Itoh, Y. Motohashi, K. Umezu, and M. Takatsuka, “Time-varying data visualization using clustered heatmap and dual scatterplot,” *The 30th Annual Conference of the Japanese Society for Artificial Intelligence, 2016*, pp. 1E4–4in2, 2016,（熊谷沙津希，伊藤貴之，本橋洋介，梅津圭介，高塚正浩，“クラスタリングとヒートマップによる高次元データ可視化” 2016 年度人工知能学会全国大会（第 30 回）.）．

[14] M. Sano, H. Masuda, K. Yamada, and T. Fukuhara, “Motivating track & field athletes by visualizing training drills and records: Extraction and visualization of activities of athletes from blog articles,” *The 30th Annual Conference of the Japanese Society for Artificial Intelligence, 2016*, pp. 1D2–1in2, 2016,（佐野正和，増田英孝，山田剛一，福原知宏，“陸上競技ブログからの活動記録抽出と可視化” 2016 年度人工知能学会全国大会（第 30 回）.）．