International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:9, 2017

Linguistic Summarization of Structured Patent Data

E. Y. Igde, S. Aydogan, F. E. Boran, D. Akay

***Abstract*—**Patent data have an increasingly important role in economic growth, innovation, technical advantages and business strategies and even in countries competitions. Analyzing of patent data is crucial since patents cover large part of all technological information of the world. In this paper, we have used the linguistic summarization technique to prove the validity of the hypotheses related to patent data stated in the literature.

***Keywords*—**Data mining, fuzzy sets, linguistic summarization, patent data

I. Introduction

HAVING an increasingly important role both in economic growth and innovation productivity, patent documents offer technical and legal information of the future developments. These documents are intensely used in industrial activities for achieving competitive advantages and in academic researches for analyzing the embedded data. The patent data seem to be major indicator of developments by helping people as indicators about how to make sense of substantial technological content, crucial tips for current and future technologic developments and forecasting technology trends [1]-[3], R&D management strategies [4]-[6], new business strategies [7], competitor monitoring and innovation abilities [8], [9].

Patent documents' data are grouped under two main categories. The first is the unstructured data, consisting of the patent documents typically texts, links, emails, messages, PDF files, photos, images, videos, or tags that are incomprehensible without inspecting in detail. The second is the structured data, also known as bibliographical data, easily entered, queried, stored, analyzed or ordered and consist of the information of the application such as filing or priority dates, publication dates, legal status of the application, inventor's names, IPC codes, number of claims, family members and abstract of the application.

Data mining, the nontrivial extraction of implicit, previously unknown, and potentially useful information from data, is defined as the science of extracting useful information from large data sets or databases [10]. When the subject is patent, the term "data mining" can be named as "patent mining" or "patent data mining". This term is defined as finding valuable, rare and non-repeatable assets of the patent documents [11] and defined as an assistance for patent analysts in efficiently and effectively managing huge volume of patent documents [12].

Linguistic summarization has been proposed as a data mining technique used for summarizing dataset by using linguistic quantifiers and summarizers characterized by fuzzy sets. Compared to the statistical summarization, linguistic summarization is an intelligent and human consistent summarization system [13]. For more details on the theory and applications of linguistic summarization, the reader is referred to [14], [15].

In this study, linguistic summarization is applied on patent data as a powerful way of examining and understanding uncover hidden patterns, correlations and other insights. Our purpose is to compare various hypotheses of patent data from the past to present with the objective results that will be directly obtained by linguistic summarization methods.

II. Background

To provide the necessary background to motivate and explain our approach, first of all, a comprehensive literature review study is conducted to better understanding the variabilities of the relations between patent data. Presented by OECD (Organization for Economic Co-operation and Development), most frequently utilized data by means of being indicators of technological and economic value are; number of patent citations, claims, different IPC codes, inventors and patent family members [16]. Considering the existing researches on the literature from the past to present, the hypotheses regarding to patent data relations are examined. Based on the literature review, researched hypotheses (H1 to H5), the relations between patent indicators, given below are to be further verified by linguistic summarization.

H1. **Q** Patents with high number of citation have high number of different IPC codes [17].

H2. **Q** Patents with high number of claims are associated with high number of different IPC codes [18].

H3. **Q** Patents with high number of different IPC codes may means to have both high number of citations and high number of patent family members.

H4. **Q** Patents with high number of inventors are associated with high patent family members [19].

H5. **Q** Patents have positive relations between number of family members and number of different IPC codes [5].

Here, **Q** is a linguistic quantifier (most, about half and few) labeled with a fuzzy set. Q is characterized in the following section.

III. Linguistic Summarization

Linguistic summarization generates natural language statements from large datasets using fuzzy sets. A fuzzy subset

E. Y. Igde is with the Turkish Patent and Trademark Office, Ankara, Turkey (phone: 0090 312 3031636; fax: 0090 312303-1173; e-mail: esen.igde@turkpatent.gov.tr).

S. Aydoğan, F. E. Boran, and D. Akay are with the Gazi University, Faculty of Engineering, Department of Industrial Engineering, Ankara, Turkey (e-mail: senaaydogan@gazi.edu.tr, emreboran@gazi.edu.tr, diyar@gazi.edu.tr).

on X, denoted by A, is defined as $A=\{\langle x,\mu_A(x)\rangle \mid x\in X\}$ where $\mu_A(x)$ is the membership degree of x. Y is defined as a set of objects $Y=\{y_1,y_2,y_3,...,y_M\}$, V is defined as a set of attributes $V=\{v_1,v_2,v_3,...,v_K\}$ and $X_k(k=1,2,...,\mathrm{K})$ is the domain of v_k. Then $v_k^m \equiv v_k(y_m)\in X_k$ is the value of the k^{th} attribute for the m^{th} object. Two forms of summary have generally used in linguistic summarization studies that are based on the fuzzy quantifiers, proposed by [20]. First of these summary forms called as type-I quantified sentence is employed in the form of *"Q Y's are / have S*$[T_1]$*"*. Here, Q is the linguistic quantifier labeled with a fuzzy set (e.g., few, very few, most, very most etc.), Y is the set of objects, S is the summarizer labeled with a fuzzy set. T_1 is the degree of truth describing how much data support the summary.

There are two basic types of fuzzy linguistic quantifiers as absolute and relative quantifiers. Fig. 1 depicts the fuzzy linguistic quantifiers.

The degree of truth for type-I quantified sentences is defined as in (1) in which $R=M$ for relative quantifiers such as "most", $R=1$ for absolute quantifiers such as "about three". The degree of truth for the type-II summaries is expressed as given in (1).

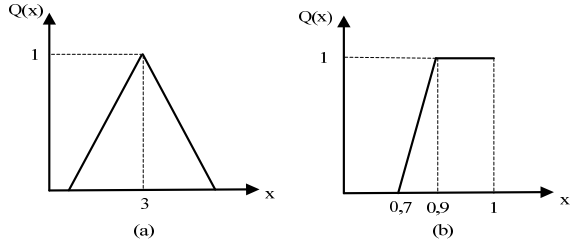

Fig. 1 Types of quantifiers: (a) "about 3" absolute quantifier (b) "most" relative quantifier

$$T_1=\mu_Q\left(\frac{\sum_{m=1}^{M}\mu_s\left(v^m\right)}{R}\right) \tag{1}$$

Type-II quantified sentence, the second summary form is given in the form of *"Q* S_g *Y's are / have S*$[T_1]$*"*. S_g is a pre-summarizer labeled with fuzzy set. T_1 is the degree of truth defined as in (2).

$$T_1=\mu_Q\left(\frac{\sum_{m=1}^{M}\min\left(\mu_s(v^m),\mu_{S_g}(v_g^m)\right)}{\sum_{m=1}^{M}\mu_{S_g}(v_g^m)}\right) \tag{2}$$

Apart from above scalar cardinality based methods for computing the degree of truth [13], [20]-[23], there are also fuzzy cardinality based methods.

GD, a fuzzy cardinality method, is used to measure the compatibility between the ED type fuzzy cardinality and a linguistic quantifier [24]. One of the commonly used methods in type-I quantified sentences $GD_Q(S)$ method is a generalized case of OWA (Ordered Weighted Averaging)-based method proposed by Yager [13]. For type-II linguistic summaries, $\Gamma\left(S\left|w_g(S_g)\right.\right)=\{\alpha_1,\alpha_2,...,\alpha_m\}$ is a set of union of α levels of $\Gamma\left(S\left|w_g(S_g)\right.\right)=\left(S\cap w_g(S_g)\right)\cup\Gamma\left(w_g(S_g)\right)$ and it holds $1=a_1>a_2>...>a_m>a_{m+1}=0$. $\left(w_g(S_g)\right)$ is defined as a normal fuzzy set. For a not normal fuzzy set, normalization should be performed. The same factor used in the normalization of $\left(w_g(S_g)\right)$ also should be used for $\left(S\cap w_g(S_g)\right)$. The degree of truth for the GD method for type-II summaries is expressed as given in (3).

$$T_{GD_Q}\left(S\left|w_g(S_g)\right.\right)=\sum_{\alpha_i\in\Gamma\left(S\left|w_g(S_g)\right.\right)}(\alpha_i-\alpha_{i+1})\times Q\left(\frac{\left|\left(S\cap w_g(S_g)\right)_{\alpha_i}\right|}{\left|w_g(S_g)_{\alpha_i}\right|}\right) \tag{3}$$

References [21] and [25] claimed the truth degree is insufficient for evaluating type-I and type-II linguistic summaries. Other validity indicators such as, the degree of imprecision, degree of appropriateness, degree of covering and degree of summary are proposed as relevant further attempt by [26].

The degree of imprecision (T_2) is a very intuitive criterion, describes how imprecise the summarizer used in a summary. (e.g., on almost all winter days the temperature is rather cold). It is defined as given in (4).

$$T_2=1-\sqrt[k]{\prod_{k=1,2,...K}in(S_k)} \tag{4}$$

where in is the degree of fuzziness expressed as:

$$in(S_k)=\frac{card\left\{v_k\in X_k:\mu_{S_k}(v_k)>0\right\}}{cardX_k} \tag{5}$$

where card is the scalar cardinality denotes the cardinality of the corresponding set.

The degree of covering (T_3) is a quality measure, expressing how many objects in $\left(w_g(S_g)\right)$ is also covered by a particular summary S, and it is defined by (6).

$$T_3=\frac{\sum_{m=1}^{M}t_m}{\sum_{m=1}^{M}h_m} \tag{6}$$

where t_m and h_m are defined as given in (7) and (8).

$$t_m = \begin{cases} 1, & \mu_S(v_g^m) > 0 \text{ ve } \mu_{w_g}(v_g^m) > 0 \\ 0, & d.d. \end{cases} \quad (7)$$

$$h_m = \begin{cases} 1, & \mu_{w_g}(v_g^m) > 0 \\ 0, & d.d. \end{cases} \quad (8)$$

The degree of appropriateness (T_4) is a quality measure that determines whether a summary is interesting or not. First, the summary is partitioned into *K* summaries (each of the partitioned summaries includes one summarizer). Then, it is computed as how many objects are in the partitioned summary by the given (9).

$$T_4 = \left| \prod_{k=1}^{K} r_k - T_3 \right| \quad (9)$$

where r_k defined as given in (10).

$$r_k = \frac{\sum_{m=1}^{M} t_{k,m}}{M} \quad k = 1, 2, ..., K \quad (10)$$

IV. Application: Linguistic Summarization of Structured Patent Data

Data, utilized in this research are, a part of European patent data including IPC (International Patent Classification) codes of Information and Communications Technology (ICT) sector published by OECD and structured from the online database (PATSTAT-Patent Statistical Database of European Patent Office) by February 2017 [27].

TABLE I
Attributes and Definitions of Patent Data

Attribute	Definition
Application ID	Technical unique identifier of the application without any business meaning.
Number of citations	Count of citations made for an application.
Number of claims	The indicator provides the count of claims.
Number of different IPC codes	Count of different IPC class symbol. (Four digits IPC codes)
Number of inventors	Count of inventors of an application.

In our study of the linguistic summarization of patent data, we empirically analyze a large random sample of 20,183 European patents. Noise, irrelevant and the insignificant zero valued contents are filtered out. We fuzzified the each pre-processed attributes in patent dataset with the modified versions of the fuzzy c-means algorithm [28]. Consciously, the pre-processed patent data consist of 17,919 data rows. The used structured patent data are composed of five attributes; number of citations, number of different IPC codes, number of inventors and number of family members is given in Table I.

A MATLAB code has been developed to generate, evaluate, and rank linguistic summaries. As a result of MATLAB runs with respect to the researched hypotheses, linguistic summaries and their associated degree of truths are obtained for three quantifiers (most, about half, few). The "most", "about half" and "few" quantifiers used in the linguistic summary structures are shown in Fig. 2.

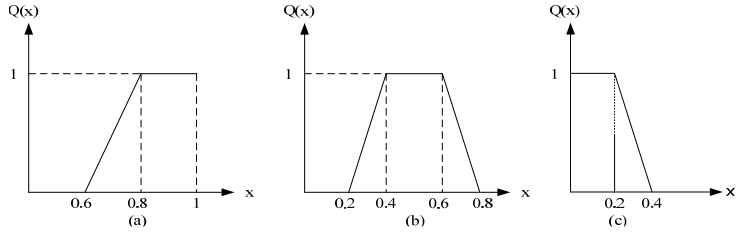

Fig. 2 The quantifiers (a) most, (b) about half, (c) few

Table II presents our linguistic summarization results. The results indicate that the results of fuzzy linguistic summarization of patent data are mostly compatible with the hypotheses. However, fuzzy linguistic summarization of patent data some gives some differences when the fuzzy quantifier is “all” or “about half”. The linguistic sentences formed with the fuzzy quantifier "about half" is evaluated as a positive result.

The only result can be referred as incompatible is the hypotheses of H4. When the source of the H4 hypothesis is examined, it is indicated that “patent inventors residing at least in two different countries” means the patent is shared between the countries. The data of inventors is not separated on the basis of the residing countries in our study so we interpreted the result as the reason why the degree of truth is low. The degree of truth of H2 hypotheses is not considered as a negative result since it was based on assumptions of “claims are drafted optimally” as noted in the author's work [18]. However, the degree of appropriateness (T_4) has the lowest value for hypotheses H2 that means the summary is not interesting. Not surprisingly, frequently observed in the literature, hypotheses H5 has the highest value for the degree of truth and also the degree of covering and appropriateness. It is obviously seen the patent data can be evaluated more

objectively and directly without any interpretation or assumption by using linguistic summarization technique.

TABLE II
TRUTH DEGREE OF THE LINGUISTIC SUMMARIES WITH THE FUZZY QUANTIFIERS "MOST, ABOUT HALF, FEW"

	TGD (Most, About half, Few)*	T2	T3	T4
H1	(0.0000, 0.9597, 0.0403)	0.7344	0.4619	0.3021
H2	(0.0000, 0.1205, 0.8795)	0.7344	0.3278	0.1679
H3	(0.0830, 0.8004, 0.1166)	0.6099	0.7686	0.5998
H4	(0.0000, 0.2899, 0.7101)	0.8097	0.5839	0.4610
H5	(1.0000, 0.0000, 0.0000)	0.7344	1.0000	0.8401

* The numbers in the parenthesis gives the degree of truth calculated with the quantifiers most, about half and few, respectively.

V. CONCLUSIONS

This study is one of the first studies in the literature that uses linguistic summarization techniques for summarizing the relations between structured patent data. The aim of this study was to evaluate the validity of the hypotheses presented in the literature with our objective and directly obtained linguistic summaries of patent data. Linguistic summaries obtained from complicated patent data are compared to the related 14 hypotheses in the literature.

The major conclusion with regard to linguistic summarization of structured patent data is that the studies conducted on the relations of structured patent data can be proven and much more detailed analysis can be implemented in specific technical fields. It is demonstrated that useful information is contained in a number of readily observable patent characteristics. With regard to positive correlations between patent data, it was found that potentially it is possible to find dozens of hidden patterns and relations that affect patents, including the patent itself, even if the applicants and inventors are not aware of. This finding strongly supports previous evidence presented by researchers.

REFERENCES

[1] Sun, T. and Y. Liu, Development of virtual reality technology research via patents data mining, in Advances in Intelligent and Soft Computing. 2012. p. 111-116.
[2] Jun, S. and D. Uhm, "Technology forecasting using frequency time series model: Bio-technology patent analysis". Journal of Modern Mathematics and Statistics. 4(3): p. 101-104, 2010.
[3] Kim, S.S. Park, and D.S. Jang, "Technology forecasting using topic-based patent analysis". Journal of Scientific and Industrial Research. 74(5): p. 265-270, 2015.
[4] Griliches, Z., "Market value, R&D, and patents". Economics Letters. 7(2): p. 183-187, 1981.
[5] Hall, B. H., G. Thoma, and S. Torrisi. The Market Value Of Patents And R&D: Evıdence From European Fırms. in Academy of Management Proceedings. 2007.
[6] Hall, B. H., A. B. Jaffe, and M. Trajtenberg, "Market value and patent citations". RAND Journal of economics: p. 16-38, 2005.
[7] Cromer, C. T., C. Dibrell, and J. B. Craig, "A study of Schumpterian (radical) vs. Kirznerian (incremental) innovations in knowledge intensive industries". Journal of strategic innovation and sustainability. 7(1): p. 28, 2011.
[8] Chen, A. and R. Chen, "Design Patent Map: An Innovative Measure for Corporative Design Strategies". Engineering Management Journal. 19(3): p. 14-29, 2015.
[9] Bessen, J. and E. Maskin, "Sequential innovation, patents, and imitation". The RAND Journal of Economics. 40(4): p. 611-635, 2009.
[10] Hand, D. J., H. Mannila, and P. Smyth, Principles of data mining. 2001.
[11] Kasravi, K. and M. Risov. Patent Mining - Discovery of Business Value from Patent Repositor ies. in System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on. 2007.
[12] Zhang, L., L. Li, and T. Li, "Patent mining: A survey". ACM SIGKDD Explorations Newsletter. 16(2): p. 1-19, 2015.
[13] Yager, R. R., "A new approach to the summarization of data". Information Sciences. 28(1): p. 69-86, 1982.
[14] Boran, F. E., D. Akay, and R. R. Yager, "An overview of methods for linguistic summarization with fuzzy sets". Expert Systems with Applications. 61: p. 356-377, 2016.
[15] Delgado, M., et al., "Fuzzy quantification: a state of the art". Fuzzy Sets and Systems. 242: p. 1-30, 2014.
[16] Squicciarini, M., D. H., and C. C., OECD-Measuring Patent Quality. 2013.
[17] Huang, M.-H., L.-Y. Chiang, and D.-Z. Chen, "Constructing a patent citation map using bibliographic coupling: A study of Taiwan's high-tech companies". Scientometrics. 58(3): p. 489-506, 2003.
[18] Tong, X. and J. D. Frame, "Measuring national technological performance with patent claims data". Research Policy. 23(2): p. 133-141, 1994.
[19] Dominique, G. and B.v.P.d.l. Potterie, The internationalisation of technology analysed with patent data. 1999.
[20] Zadeh, L. A., "A computational approach to fuzzy quantifiers in natural languages". Computers & Mathematics with applications. 9(1): p. 149-184, 1983.
[21] Kacprzyk, J. and R. R. Yager, "Linguistic Summaries of Data Using Fuzzy Logic". International Journal of General Systems. 30(2): p. 133-154, 2001.
[22] Yager, R. R., "Database discovery using fuzzy sets". International Journal of Intelligent Systems. 11(9): p. 691-712, 1996.
[23] Bosc, P. and L. Lietard, "On the comparison of the Sugeno and the Choquet fuzzy integrals for the evaluation of quantified statements". Proceedings of EUFIT. 95: p. 709-716, 1995.
[24] Delgado, M., D. Sánchez, and M. A. Vila, "Fuzzy cardinality based evaluation of quantified sentences". International Journal of Approximate Reasoning. 23(1): p. 23-66, 2000.
[25] Wu, D. R. and J. M. Mendel, "Linguistic Summarization Using IF-THEN Rules and Interval Type-2 Fuzzy Sets". Ieee Transactions on Fuzzy Systems. 19(1): p. 136-151, 2011.
[26] Kacprzyk, J. An interactive fuzzy logic approach to linguistic data summaries. in Fuzzy Information Processing Society, 1999. NAFIPS. 18th International Conference of the North American. 1999.
[27] Office, E. P. Patstat. 2017; Available from: https://www.epo.org/searching-for-patents/business/patstat.html#tab1.
[28] Dunn, J. C., "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters". 1973.

Igde Esen Yıldız is a Junior Patent Examiner in Turkish Patent and Trademark Office. She received the degree of B.Sc from Çukurova University, Adana in 2009. Her theme of study has involved patent analysis and new technology development methodology.

Aydoğan Sena received the B.Sc. and M.Sc. degrees in Industrial Engineering from Gazi University, Ankara, Turkey, in 2012 and 2015 respectively. She is currently an Research Assistant with the Department of Industrial Engineering, Gazi University. Her current research interests include fuzzy sets and systems, and decision making.

Boran Fatih Emre received the B.Sc., M.Sc., and Ph.D. degrees in Industrial Engineering from Gazi University, Ankara, Turkey, in 2007, 2009, and 2013, respectively. He is currently an Associate Professor with the Department of Industrial Engineering, Gazi University. His current research interests include fuzzy sets and systems, and decision making.

Akay Diyar received the B.Sc., M.Sc., and Ph.D. degrees in Industrial Engineering from Gazi University, Ankara, Turkey, in 2001, 2003, and 2006, respectively. Between 2007 and 2009, he was a Post-Doctoral Researcher with the University of Leeds, Leeds, U.K. He is currently an Associate Professor with the Department of Industrial Engineering, Gazi University. His current research interests include fuzzy sets and systems, and affective design.