3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11645

Extracting usual service prices from public contracts

Tomáš Bruckner, Filip Vencovský
Department of Information Technologies, University of Economics Prague, Czech republic.

Abstract

The paper describes a project of automatic selection, scraping, and full-text analysis of contracts in the area of IT and Information Systems. The purpose of the project was to extract manday prices and build the list of usual manday prices for particular roles that are stated in the contracts. The list aims to provide a foundation for sizing of new IT solutions before the public tender for an association of major state institutions of the Czech Republic. The result of the research is the list of usual prices for the specified roles, including blended rate, based on median and interval between quartiles, all with demonstrable links to origin contracts. The discussion states additional social factors to be considered when interpreting and using the resulting list, like the subjective influence of validators, tendency for generalization, or defensive attitude of affected vendors.

Keywords: *usual price; contracting; full text analytics; information technology, big data.*

1. Introduction

Usual prices are used in businesses to set the estimated value of procurement. Especially IT or Information Systems contracts tendered by public institutions are sensitive for correct sizing (Ochrana & Pavel, 2013) due to the special rules of public tenders where preliminary negotiations for the purpose of finding expected price can be challenged by competitors who were not addressed. Public institutions struggle with the need of usual price determination before any public tender. Therefore, thirteen IT departments of major state institutions joined in an association with the purpose of tenders facilitation and assigned and sponsored this research.

The usual prices are generally obtained by a counsellor, who takes few demonstrably negotiated contracts and proves prices negotiated in a given time and place, as stated in (Act. No. 526/1990). In IT and Information Systems contracts, such findings can vary greatly and hence are not very useful for contract sizing. Due to that, the sponsor decided to order research of broad usual prices overview, based on all contracts demonstrably negotiated by public institutions in the Czech Republic, which are available in the public contract register (Act No. 340/2015).

The research problem is a complex application of technology and business topics. The technology consists of web scraping, unstructured data analysis, information extraction, information quality and data validation, business intelligence and data visualization. The business part consists of understanding the content and principles of contracting and pricing in IT and Information Systems. The primary source of the data is a public register of contracts, available on the internet, which contains millions of contracts. The problem is to restrict the amount to those contracts, which are relevant to IT and Information Systems, read the contracts, identify the place and time of validity, identify price details, extract the price details, classify the pricing data, decide the right usual price and visualize and present the data in the form beneficial for the sponsor institution employees on a daily basis.

The level of correctness has to be high because of the result is observed and verified by the potential vendors of the future contracts. The vendors felt offended in the preliminary stage of the research; arguing the projects dictates the contract prices and restricts freedom of price setting and giving the sponsors an overview about different prices the same vendor contracted with different state institutions. Therefore, the communication of the research and its results must be very accurate and cautious.

The prices of contracts in IT and Information Systems are complex pricing systems assembled of unit prices. The unit prices could be the prices of services, such as licenses, access to applications, computing time, or products like hardware. This paper focuses on a specific unit price - a manday price of work of a specialist.

2. Methodology

We decided the main measure of the result is correctness, where every price included in the final result should be traceable to the original contract. Thus, every possible protest against the result could be argued easily. The proposed list of the usual prices shows typical roles of specialists in IT or Information Systems area. The usual price was calculated for each of the roles as an interval based on a set of traceable data extracted from the contracts. The additional value for users of the list was provided by deeper classification of each role using tags. Hence, the price can be estimated more precisely for various conditions. The list of usual prices was accompanied by a benchmark of prices negotiated by the sponsor institutions. The initial role set, a role codebook, was interviewed with future users of the list within sponsor institutions.

As the prices may change in time, the list should be repeatedly assembled in a certain interval. We set the interval to six months in which the two years of historical data will be evaluated. Due to the repeatability, the research method should be as automated as possible and independent on individual decisions of the researches of future runs but should enable partial innovations of the process. We elaborated the documented research process, consisting of sub-processes: Preparation, Contract acquisition, Prices extraction and validation, Usual price determination, Handover and acceptation.

In the preparation phase, besides the infrastructure preparation (the technological architecture composed for the research is the subject of another paper (Bruckner, 2019)), a list of organizations operating in particular business areas concerning IT and Information Systems is created. The reason of the list is to exclude contracts, which are surely not related to the area in question. For that purpose, the NACE activity classification (Nomenclature statistique des activités économiques dans la Communauté européenne) from the Public Business registry is taken into account.

The contract acquisition phase is performed as a combination of scraping the frontend of the public register of contracts and API access to the register. Only contracts of subjects listed in the previous step by NACE are scraped. The scraping resulted in the set of contracts that were downloaded in the next step. Each downloaded contract is transformed into plain text, whereas the original form is preserved for a possible visual check. In some cases, OCR over an image must be performed. All contracts in plain text are reverse-indexed, lemmatized, stemmed and stored in the search engine (Zobel & Moffat, 2006)□. As the search engine, we employed Elasticsearch (“Elastic Enterprise Search,” 2020), a software that uses Apache Lucene search library. Every half of a year a new run is indexed into a separate database. Every contract is identified by a specific number, new version of the same contract rewrites older version.

The prices extraction phase consists of document indexing, search, and processing. As the contracts were indexed, we were able to create sets of queries for the selected typical roles of specialists in IT or Information Systems area to identify contract documents, in which a manday price for those roles probably occurs. The results of the queries were experimentally verified by experts.

The queries were executed on the search engine. The documents were processed by price extraction scripts in the next step. The price extraction is realized by a set of regular expressions (Brauer, Rieger, Mocan, & Barczynski, 2011; Mooney & Bunescu, 2005) which match several possible ways to describe a manday price in the text of the contract and which most likely do not match other data than prices, e.g. quantity of mandays etc. The regular expressions were manually verified on a subset of data and were a subject of fine-tuning during the price extraction.

In the validation phase, the group of experts verified the correctness of extraction of every price data. The extreme or unusual data were validated by two experts at least. Data for validation were sorted ordered according to their relevance based on the TF-IDF measure (Salton & Buckley, 1988) in a large table, with several contract metadata, and submitted to a group of experts for validation.

In this paper, we focus on the method of determination of the resulting list of usual prices. Thus, the core activity took place in the usual price determination phase. The handover and acceptation phases are not described in this paper.

We define the usual price as a price for which, in given time and given place, was provided (sold) a similar service or product. The given time, we understand as the period of two following years for which we gathered the data. The place, in this case, is the territory of the Czech Republic.

By a similar service, we understand a provision of a service priced by a rate for a time interval (an hour or a day) when providing service in the context of information technologies or information systems. For the classification of a particular service to one of the predefined roles, the description of the service in the contract is determinative. We reserved the right to classify hazily or shortly defined service descriptions inaccurately.

All price data were normalized to a manday price, a price for eight hours of work, in local CZK currency, VAT exclusive. If the currency is not CZK, the price is converted by Czech CEB conversion rate valid at the day of signing the contract.

Due to the variability, we do not determine the usual price as a price of single services but as a statistical calculation from a set of single particular services in given time and place for every class of typical role. From the set, we compute the following descriptive statistics: lower quartile, median, upper quartile, minimum, average, and maximum.

For the determination of the usual price, we suggest not to take into account the extreme values. Thus, we consider the usual price as the median. Moreover, we define a usual price spread as the spread between quartiles, or the interval, which do not include a quarter of lowest and a quarter of highest prices within the given role set.

Due to the method of data acquisition, prices for the same role from one contract were obtained multiple times. There were cases in which the stated price differs for the same roles; usually due to the different time of negotiation. Therefore, if the set includes different prices for the same role classified by the role codebook and tag description from the same contract, such prices were included only once, namely as an average.

If there were stated different price data for different roles in one contract, all of them were included in the data table separately. The same rule was applied in case of different specification of the role was detected, e.g. different seniority or request response time. The different specification data were included in the data table with the use of corresponding tags.

Considering the existence of various tags, we included tag consolidation activity into the usual price determination phase. In the time of extraction and validation, we created a semantic mapping of different tags. This mapping was used to join small sets of differently tagged price data to bigger, more relevant sets. In this step, the size of the set and the semantic generalization of the tags were considered. The consolidation was done by regular expressions with several variants of tag generalization.

The final output was a set of tables: summary table of all characteristics for roles, summary histograms for particular roles, and summary box plot chart in the same scale for roles. Furthermore, for every role was created a separate table where, apart from summary values for the role, values for specific subsets defined by consolidated tags were presented.

3. Results

We completed the list according to the described method five times, for the first time in half of 2017, for the last time in the half of 2019. The last run contained the contract data from May 2017 to May 2019. The method has undergone some slight innovation on the way. The method and the results presented in this paper are from the last run.

The complete set of all contracts in the public contract register counted millions of documents. After the NACE restrictions, we scraped amount of 566,315 contracts. After the price extraction step, we got a set of 10,051 candidate price data items. After manual validation and tags consolidation, we finished data preparation with 6,408 price items from 2,336 unique contracts from 776 public contracting authorities and 884 vendors.

Table 1. Summary of usual prices for roles.

Role	Lower quart	Median	Upper quart	Min.	Avg.	Max.	Con-tracts	Ven-dors	Pri-ces
administrator	5,920	7,600	10,560	1,200	8,567	24,000	213	133	357
analyst	8,000	10,400	13,580	1,440	10,701	24,000	341	149	402
architect	7,422	11,920	12,800	1,500	10,906	42,368	140	55	200
auditor	9,600	11,400	14,830	6,400	12,079	20,400	18	18	19
blended_rate	7,376	9,600	12,000	800	10,099	47,120	559	285	671
support	6,000	8,000	12,000	8	9,364	62,400	521	283	757
coder	7,260	9,600	11,940	1,920	9,697	30,560	553	242	652
routine_jobs	4,720	8,000	10,400	1,512	8,297	24,000	174	90	238
specialist	8,000	10,000	12,000	1,040	10,689	42,368	849	336	1,650
lecturer	6,400	9,600	12,000	1,800	9,952	43,200	360	167	433
technic	4,590	6,400	8,650	40	7,122	31,680	380	289	612
tester	5,680	7,912	9,975	1,232	8,038	20,584	105	62	123
project_manager	8,160	12,000	14,400	1,300	11,852	28,000	236	104	294

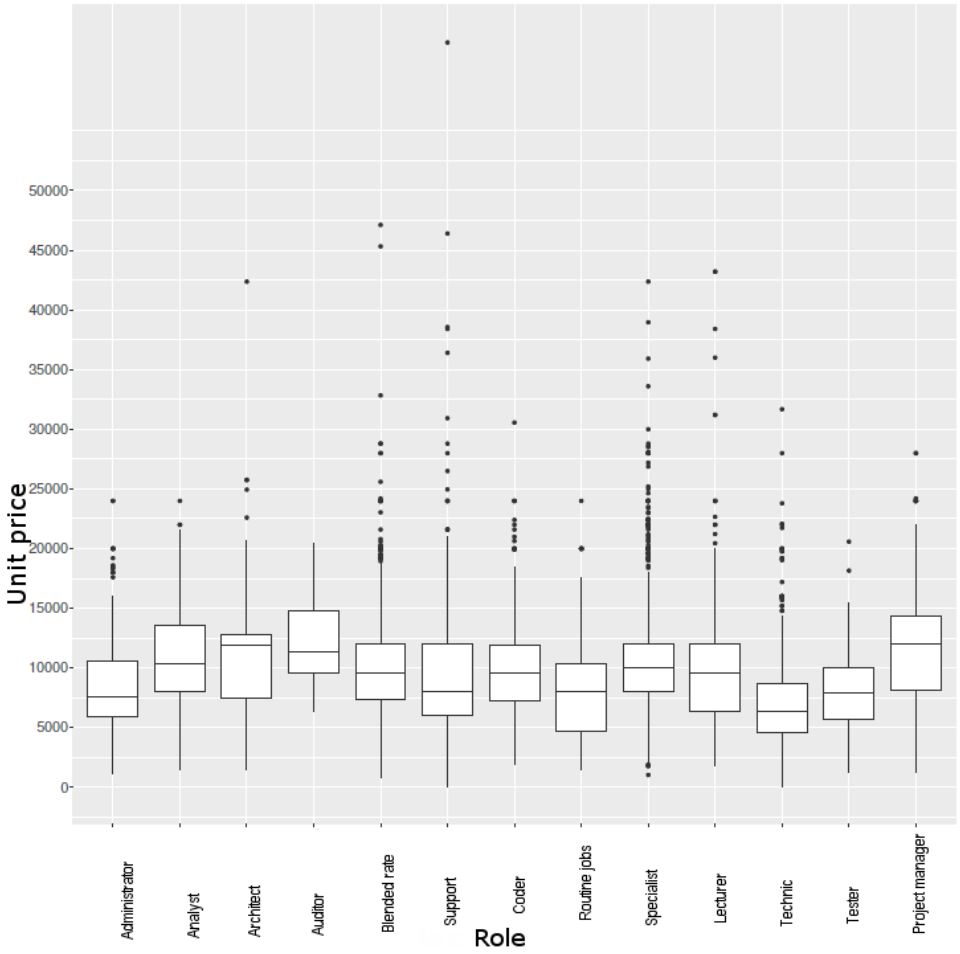

Figure 1. Summary of usual prices for roles.

The found usual prices are shown in Table 1. The graphical representation is depicted in Figure 1. As we can see, the result prices are slightly different for focal roles: The most expensive roles were project manager, architect and auditor, the least expensive was a technic. The role with the highest price variance was a support, which also brought the price with the absolute highest value.

A specialist was the role with the highest occurrence in the set. The area of specialization causes the difference in values within the role. An auditor (understood the IT or Information systems auditor) was the role with the least occurrence. The smallest price with eight Czech crowns per manday was stated in a contract with a high volume of other products and services, and hence probably the price was compensated in this case.

The detailed results, the dataset of all price data with links to original contracts, and the histograms and tables for selected roles are beyond this paper.

4. Discussion / Conclusion

While the numerical results of the research looked obvious and clear, there were aspects which should be considered when interpreting the results.

Majority of the accessed contracts were complex contracts which cover more than manday prices. Thus, the prices might be mutually compensated, which worsens objectivity of the result. Furthermore, the object of a contract is usually described in detail while the service description of the contract for the roles was often described very vaguely. On account of that and since the unit price was not cleansed from other pricing factors, it is not possible to make any conclusions about the real price for the selected role on the market.

Also, during the validation, the incorrectly extracted price data was corrected by the experts which was affected by a subjective view of the expert. The unclear cases could be (and was) classified by different experts differently.

Some contracts defined a unit manday price, although the roles were not named or further specified. For this reason, we defined a special category called Blended rate where the unit price is stated, but the role is indefinite.

Moreover, only contracts in the public sector are published in the registry, which was our only input the contracts between private subjects are confidential and not available for the research. We expected the prices in the public sector to be higher than in the private area. However, based on a discussion with the validation experts, the found unit prices tend to be rather lower than unit prices in private contracts while the overall price volume of the contracts tends to be higher. The reason could be the criteria of the selection, where the unit price is taken into account.

All those aspects decreased the possibility of generalization of the found price level to the whole market. The results only state which prices were contracted, and are very likely biased by the chosen details of the method. At least, the results could be useful for the statement of grounded theory for future research.

The risk is that the public sector workers may interpret the extracted usual price from the results as the price which is “desirable” and “normal” and use it not only for preliminary contract sizing for the tender but also for the negotiation with the vendors. This is accelerated by the government of the Czech Republic, which stated the list of usual prices as a mandatory consulting material for all state institutions public tenders.

References

Act No. 340/2015 Coll., on special conditions for the effectiveness of some contracts. Czech Republic.

Act. No. 536/1990 Coll., on prices. Czech Repiblic.

Brauer, F., Rieger, R., Mocan, A., & Barczynski, W. M. (2011). Enabling information extraction by inference of regular expressions from sample entities. In International Conference on Information and Knowledge Management, Proceedings (pp. 1285–1294). New York, New York, USA: ACM Press. https://doi.org/10.1145/2063576.2063763

Bruckner, T. (2019). Design of the technological architecture for PUMPIT project. Journal of Systems Integration 10(2) (2019) 34-40

Elastic Enterprise Search. (2020). Retrieved February 1, 2020, from https://www.elastic.co/enterprise-search

Mooney, R. J., & Bunescu, R. (2005). Mining knowledge from text using information extraction. ACM SIGKDD Explorations Newsletter, 7(1), 3–10. https://doi.org/10.1145/1089815.1089817

Ochrana, F., & Pavel, J. (2013). Analysis of the Impact of Transparency, Corruption, Openness in Competition and Tender Procedures on Public Procurement in the Czech Republic. Central European Journal of Public Policy, 7(2), 114–134.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Zobel, J., & Moffat, A. (2006). Inverted files for text search engines. ACM Computing Surveys, 38(2), 6-es. https://doi.org/10.1145/1132956.1132959