Case-Based Appraisal of Internet Domains

Sebastian Dieterle and Ralph Bergmann

Business Information Systems II
University of Trier
54286 Trier, Germany
{diet5101,bergmann}@uni-trier.de
www.wi2.uni-trier.de

Abstract. The increased economic importance of the Internet in all branches of industry has turned Internet domain names into "virtual properties" for electronic commerce. However, the appropriate appraisal of a domain name is a great challenge for trading partners. To address this problem, we propose a new approach derived from the sales comparison approach largely applied in the real estate market, which is implemented using case-based reasoning. The required case base that contains the experience in previous domain trade transactions is automatically extracted from the Internet. Based on carefully selected features derived from the domain name, cases are retrieved using a knowledge-intensive similarity measure. Further, case adaptation is performed to adjust the sales value of a previous domain transaction with respect to the difference to the target domain being assessed. The proposed method is implemented as a prototype Internet Domain Name Appraisal Tool (IDNAT). The performed empirical evaluation using 4,231 cases from the .de domain clearly demonstrates the feasibility of the proposed approach.

Keywords: CBR Application, Appraisal, Case Representation and Mining, Similarity, Adaptation.

1 Introduction

When a new website is to be launched, the selection of the domain name is of crucial importance, particularly due to the increasing shortage in namespaces. Companies and organizations are faced with the difficult task of determining what an appropriate and easy to remember web address is worth. Hence, the appropriate appraisal of an Internet domain such as `hotel.de` is a very challenging task for the partners involved in domain trading transactions. Currently, several domain appraisal services in the Internet offer related services, but existing valuation formulas for domain names are not able to handle the complexity and ambiguity of the natural language used in a domain name. Overall, the valuation process is still not completely understood. Prices achieved in domain trade depend not only on the nature of the respective domain names, but also to some degree on the specific situation of the trading partners such as shortage of money, lack of market knowledge, personal significance of a domain name and the negotiation skills of the market players [1].

B. Díaz Agudo and I. Watson (Eds.): ICCBR 2012, LNCS 7466, pp. 47–61, 2012.
© Springer-Verlag Berlin Heidelberg 2012

To address the problem of domain name appraisal, we propose a new approach derived from the sales comparison approach largely applied in the real estate market [2, 3, 4]. It compares recently sold similar properties (in the local neighborhood) to the target property and adjusts their sales price to compensate for the differences. This method is related to case-based reasoning (CBR) and hence, has already been implemented, thus using CBR for real estate appraisal [5, 6, 7]. However, to our knowledge, CBR has not yet been applied to the appraisal of Internet domains. Therefore, we adapt and elaborate this approach for the appraisal of Internet domains and demonstrate its feasibility. In this approach, the required case base that contains the experience in previous domain trade transactions is of core importance. Publicly available domain sales data, for example, from Namebio [8] or United-Domains AG [9] is an important source from which cases can be extracted. However, the pure strings representing the domain name are not immediately appropriate to assess the similarity of two domains.

It is important that more informative features are derived. This makes it possible to better characterize the domain name with respect to its sales value. We address this problem by deriving an elaborated case representation of domain name transactions containing relevant attributes that can either be calculated from the domain name by a generic function or which can be extracted by use of existing Internet services such as Google, TEOMA [10], or Open Directory Project (ODP) [11]. Hence, we follow the strategy to perform the domain name appraisal based on case-based reasoning on the Web.

The following section describes related work on domain name appraisal and on CBR for appraisal tasks. Section 3 then presents the details of the domain name appraisal approach and the implemented Internet Domain Name Appraisal Tool (IDNAT). The results of an empirical evaluation using an extracted case base of 4,231 domain name transactions from the .de domain are presented in Section 4 before final conclusions are drawn in Section 5.

2 Related Work

2.1 Internet Domain Appraisal

A domain name such as `domain.de` consists of two parts: the freely selectable Second Level Domain (SLD) before the dot and the respective Top Level Domain (TLD) after the dot. The sales value of a domain depends on various aspects. The so-called RICK Formula [12] was the first attempt to make a proposal on the value of a domain. It is well-known in literature and generally applicable, but it is very informal. It claims that a good domain should be free of legal *risks*, i.e., it should not contain terms that refer to other brands or label rights. Furthermore, it has to show a strong *image* TLD, e.g. `.com` for the international market or `.de` for the German-speaking region. The third aspect is the *commercial potential* of a domain, i.e., an assessment as to whether it is suitable for the operation of a profitable business model. The last

and most important point during the selection of a domain is the “golden rule”, which states that the domain name should be *kept* as *short* as possible, i.e. the length is a crucial issue.

The fundamentally different SCHARF Model [13] is a factor model that allocates score points for 42 different characteristics of a domain. They are aggregated and with the help of a conversion table, the total number of points is turned into a concrete monetary value. In addition, certain scoring penalties are applied. For example, the value of the domain is decreased if it contains a number, a hyphen or a special character, because this can result in various ways of spelling and hence, spellings mistakes are more likely.

The recently proposed model by Sunyaev et al. [14] weights eleven numerical criteria and determines a precedence of similar domain names by means of a recursive tracking approach. For this purpose, the user initially enters search terms and by means of a Google search, further domain names are found, which in turn, are again used as search terms, etc. The domain names are placed in a precedence list, but a monetary valuation is not executed. This approach is not primarily concerned with assessing, but rather with finding existing domain names, as well as with the creation of valuable domain names with economic potential. In this context, the names that are found are supposed to thematically suit a planned project from a user's point of view. In addition, similar criteria are used as in the two previously described approaches.

Market-oriented domain appraisal approaches, such as Estibot Version 2.0 [15], have recently gained an increasing popularity. Estibot determines the sales value of a domain based on the grounds of similar sales, because these are supposed to best reflect the current market value. Transactions are only considered similar if they exactly contain one word as a substring that occurs in the domain being valued. For example, if the appraisal query is `airlinereservations.de`, all transactions which exactly have `airline` or `reservations` as a substring are displayed. With this, another important valuation criterion is added, i.e., the meaning of the domain name. This reflects the fact that a domain cannot be appraised in a manner independent of the profitability of a branch.

The approach described in this contribution consists of a combination and diversification of the existing appraisal approaches. In particular, it is also based upon earlier transaction cases, but applies a more elaborated similarity function on a larger set of comparison criteria, including structural and textual components.

2.2 Case-Based Appraisal in General

To our knowledge, CBR has not yet been applied to the appraisal of Internet domains, but it was recognized early as a technique capable of supplementing automated approaches for the appraisal of real estate with the help of market data. As an alternative to case-based appraisal, multiple linear regression [16], artificial neural networks [17] and traditional rule-based expert systems [18] have also been applied to appraisal. However, these are not discussed further here.

Gonzalez & Laureano-Ortiz [5] were the first to use CBR to value real estate in 1992. CBR corresponds most closely to the cognitive process of human appraisers. It solves problems by adjusting earlier experiences to a new situation. Since the solution consists only of one single attribute in the appraisal domain, it is very important to find the best possible solution. As a problem description, they used the size of the living space, the number of rooms, the age of the house, the date of sale and the type of heating. The adaptation rules for price adjustment are defined in advance by an expert and the weighted average of the most similar comparison object is computed [5]. Bonissone & Cheetham [6] propose an approach (called PROFIT) for valuing property ownership. In order to define the similarity between the subject entity and the comparison object, this approach combines case-based reasoning with fuzzy logic. McSherry [7] presented an approach implemented with CREST (Case-based Reasoning for ESTimation) that allows for the adjustment of comparison objects independent of the application domain, yet under the assumption of an additive value function.

The appraisal of Internet domains differs from that of real estates since linguistic aspects must be considered for Internet domains and since the geographic proximity plays no role at all. Hence, other criteria that can be automatically derived from the domain name must be determined.

3 CBR for Domain Appraisal

Our approach to case-based domain appraisal requires a case base of experience in previous domain trade transactions. The selected attributes contained in the case representation (described in Section 3.1) must enable an appropriate assessment of the similarity of two domain names and must also provide the information required for adaptation. While the case representation is usually static, the cases themselves must reflect the currently available experience in domain sales transactions on the Web and hence they will be extracted automatically from existing Internet services (see Section 3.2). The automatic extraction is essential in order to keep the case base up-to-date by integrating recent domain sales transactions.

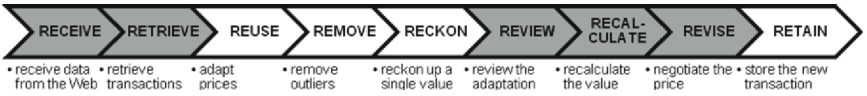

Fig. 1. Pricing a domain name using CBR

To appraise a new target domain, the user simply enters the desired domain name as a query and the appraisal process depicted in Fig. 1 is started. From the domain name, a new query case is generated by automatically deriving the relevant features from the domain name query string. We call this the *receive* step as additional data is received from the web to compute these features. In the subsequent *retrieval* phase, a set of similar cases is retrieved from the case base (currently we use the 11 most

similar cases), using a knowledge-intensive similarity measure (see Section 3.3) based on the derived features. Then, the *reuse* phase performs the adaptation of each retrieved case, i.e., of the solution attribute containing the sales value of the sales transaction in the selected case (see Section 3.4). As a result, each retrieved case now has a new sales value assigned that reflects an assessment of the value of the target domain, based on the experience of the individual transaction of the case. In order to *remove* noise from the assessment results, outliers are removed and a weighted average (case similarity is used as weight) of the remaining assessment values is computed (*reckon* step). This value is the proposed appraisal value for the target domain. Professional users may then *review* the computed solution. If necessary, they can tweak the computation process by modifying the used cases and weights (see Section 3.6), which leads to a *recalculation* of the appraisal value. In the *revise* phase, the negotiation of the sales price among the trading partners takes place, based on the appraisal value as a starting point. In case the negotiation is successful, the final sales price is then used in the *retain* phase and leads to a new case that is stored in the case base. Selected details of this approach are presented in the rest of this section.

3.1 Case Representation

The case representation (identical for case and query) was developed such that it contains those characteristics of the domain name that are relevant to determine whether two domains have a similar sales value. A traditional attribute-value representation including symbolic, numeric, textual, and multi-valued attributes was sufficient for this purpose. Altogether, 27 attributes were identified (see Table 1 for the most important attributes). They serve the definition of similarity and are used to control adaptation of the solution, as well as to enable the traceability of the cases.

First, there are basic attributes that form the foundation of every case and from which all other attributes are derived. The basic attributes are the domain name (with SLD and TLD), the date of the transaction (with year and month) and the achieved sales price. The sales price is the only solution attribute in this representation, and hence a value is not available in the query. Neither a usable similarity measure nor useful adaptation rules can be defined directly on the SLD. For this reason, the spelling separated into words and the spelling separated into word components are considered. The difference between the two kinds of spellings is, for example, that the SLD "ArtificialIntelligence" consists of two words and also of two word components. However, SLD "bookworm" is only a single complex word, consisting of two word components ("book" and "worm"). This differentiation is of importance to better discern the semantics of a domain name. On the one hand, cases are retrieved in which the same or related words appear and on the other hand, the same or related word components determine which cases are selected. Although more results can be found with the consideration of individual word components, the quality of the results is higher when complete words are used. As an additional attribute, a classification of the domain name into a set of categories is used. The categories come from a

hierarchical classification of the content domains of all web pages world-wide, as provided by the Open Directory Project (ODP) [11]. The category-value in the case representation is constructed in the form of a path in the hierarchy, for example: `World:German:Recreation:Travel:TravelReports:Europe:Spain`.

Table 1. Partial description of a query case and weights for retrieval

Attribute	Exemplary instance	Weight in retrieval
Domain name	`spanienurlaub.de` (German for "holiday in Spain")	-
SLD	spanienurlaub	-
TLD	de	-
Transaction year	2012	0.05
Length	13	0.1
Number of words	2	0.03
Number of word components	2	0.02
Category	World:German:Recreation:Travel: Travel Reports:Europe:Spain	0.15
Words	spanien urlaub	0.1
Word components	spanien urlaub	0.05
Search results	29,100	0.02
Contains hyphen	false	0.02
Number of hyphens	0	0.01
Contains special characters	false	0.01
Contains numbers	false	0.01
Domain age	1999.25	0.01
Global monthly searches	9,900	0.01
Local monthly searches	9,900	0.1
Avg. CPC in €	4.3	0.01
Daily clicks	23.79	0.01
Daily cost in €	102.33	0.05
Sales price in €	unknown	-

Furthermore, attributes that describe the length of a domain name are included. This relates to the number of individual characters, the number of words, as well as the number of word components. Other visible, linguistic characteristics of domains are differentiated by whether or not and how many hyphens or special characters such as ä, ö, ü, ß, á, à etc. or numbers exist in the SLD.

Other characteristics that have similar relevance must also be considered. This includes the domain age, i.e., the date of registration. In addition, the number of search results in search engines must be considered, i.e., the number of websites that contain the SLD in a separated spelling. Furthermore, the number of searches performed worldwide and regionally in countries related to the TLD is a characteristic that is also included, i.e., how often the search term is searched for per month. In addition, the average price per click on the advertisement for the corresponding search term, the number of daily clicks, and the total costs for advertising the search term are considered.

3.2 Data Extraction

The information regarding the selected characteristics must be calculated in an automated manner or extracted from the Internet in order to be able to automatically update the case base.

First, domain sales transactions were extracted from the Internet using existing services, such the world's largest public list from Namebio [8] or the domain sales list from United-Domains AG [9]. Double entries, obviously erroneous entries, syntactical incorrect values, as well as older values were filtered out. Then, the secondary attributes had to be derived from the extracted domain transactions. They can either result from the extraction of data from the Internet or from the combination of existing attributes. Here, a large number of specific algorithms were implemented (see Dieterle [19] for details). For example, based on a heuristic, the domain name is divided into the individual words and word components if these are not already separated by hyphens. This takes place through the assessment of how common the variant forms of spelling are by using Google queries.

Among other things, the SLD must be classified in the categories of the Open Directory Project (ODP) [11]. The ODP is the world's largest web directory, which is maintained by nearly 95,000 volunteer editors. It contains a hierarchy of more than 1,010,000 categories and may be used open source, for example, for the development of Artificial Intelligence (AI) applications. A large number of research projects [20] make use of this opportunity. During the data extraction, up to three different, compounded categories are extracted, to address the potential ambiguity of the words. Thus, a domain such as `crimefiction.de` fits in a category for crime novels, for crime series, and also into the category for computer games related to the genre of crime.

3.3 Similarity

This subsection presents the similarity measure used during the retrieval of similar cases. Please note that the goal is to define a knowledge-intensive similarity measure such that it approximates the *utility* of the case (see [23] chapter 4) for the current query. Given the fact that the sales value of a case is possibly adapted after retrieval, a case should also be considered useful to a query if its sales value can be adapted such that it is similar to the real sales value of the domain in the query.

Given the traditional attribute-value representation for cases, the well-established local-global approach for defining similarity measures is applied. As an aggregation function, the weighted sum of 18 local similarity functions is used. Table 1 illustrates how the local similarity functions are weighted. The weighting has an experimental character and is based on the following general considerations:

- attributes with a high relevance for the value of the domain should have a high weight,
- attributes whose differences can be well compensated by adaptation should have a low weight.

Different types of local similarity functions are defined, depending on the type of attribute: textual similarity functions define the similarity between two unstructured strings (e.g. the SLD), symbolic similarity functions define the similarity between symbolic values (e.g. categories), and numerical similarity functions compare numerical attributes (e.g. the length of a domain). These similarity measures are further specialized for the respective attribute (see Dieterle [19] for details).

In the following, two selected local similarity measures are presented. The first similarity measure $sim_{words}(Q, C)$ allows for the semantic comparison of words (or word components) of a domain.

$$sim_{words}(Q,C) = \frac{\sum_{q \in Q} \max(\{sim_{word}(q,c) | c \in C\}) + \sum_{c \in C} \max(\{sim_{word}(q,c) | q \in Q\})}{|Q| + |C|} \quad (1)$$

The arguments Q and C stand for a set of words contained in the query or case, respectively. The maximum similarity is derived for every individual pair of words using the similarity measure $sim_{word}(q, c)$. This measure compares two single words (or word components) q and c with each other.

$$sim_{word}(q,c) = \begin{cases} 1 & if\ q = c \\ \frac{\min(length(q), length(c))}{\max(length(q), length(c))} & if\ \exists s \in substring(q): s = c \vee \exists s \in substring(c): s = q \\ 0.8 & if\ \exists r \in relatedWord(q): r = c \vee \exists r \in relatedWord(c): r = q \\ 0 & otherwise \end{cases} \quad (2)$$

If q and c are identical, the similarity is the maximum. If one word is a sub-string of the other word, the similarity corresponds to the degree of overlap of both words. For example, “business” and “businesses” correspond with eight of ten letters (0.8); while “men” and “investment” correspond with only three of ten letters (0.3). In the case that the two words are related to one another, a reasonably high similarity value of 0.8 is assumed. The relatedness of two words is determined using the search engine TEOMA [10]. When entering a search term into TEOMA, a list of related search terms are also displayed. They are determined by TEOMA by mining frequent sequences of terms that users enter as queries. For example, the terms beach, ocean and palm trees are related to the term vacation. Alternatively, computational-linguistic approaches could have been used for semantic relatedness and similarity [21, 22], but they seem less useful for our purpose, as they are based on the occurrence of words in written documents or in dictionaries such as WordNet. Hence, they are not necessarily aligned with the natural search behavior of Internet users and the vocabulary used during a direct input of an Internet address.

The second local similarity measure that we now present is used to compare the category attribute. Here, a taxonomy is used to determine terminological similarity. Taxonomies are n-ary trees, whose nodes contain symbolic values. Unlike the taxonomy similarity measures proposed by Bergmann [23], we apply a graph-theoretical approach [24], which does not require assigning similarity values to each node of the taxonomy. Given the fact that the range of categories spans the whole world-wide spectrum of possible topics of a Web page, it is obvious that this taxonomy is extremely large (as we use the Open Directory Project it contains more than 1,010,000

categories), hence a manual assignment of similarity values is not feasible. Instead, the distance between the nodes in the taxonomy is used to define a similarity measure. The similarity between the category of a query and a case is computed by the following formula:

$$sim_{category}(q,c) = \begin{cases} 0 \; if \; countEdgesBetween(q,c) > 10 \\ 1 - 0.1 \cdot countEdgesBetween(q,c) \;\; \text{otherwise} \end{cases} \quad (3)$$

The example in Fig. 2 plots the similarity between the node "women's soccer" and the node "Soccer World Cup 2006". Since there are four edges between them, the similarity according to (3) is 0.6. This form of similarity determination is efficient because it only considers the categories from the query and the case (and their paths in the taxonomy). The remaining 1,010,000 categories do not need to be considered.

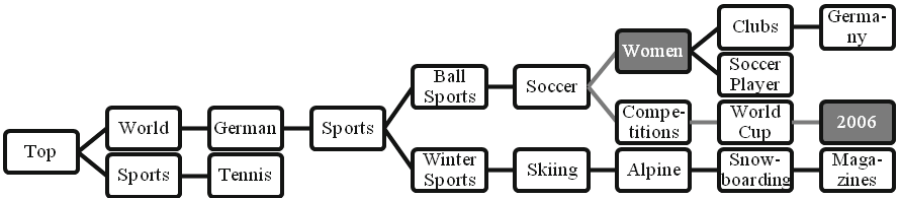

Fig. 2. Example of the distance between two cases within the taxonomy

3.4 Adaptation

The most similar cases found are not just reused unmodified, but their solutions, i.e., the sales value attribute is adapted by use of multipliers. In the area of company valuation (particularly when using the market value method), it is well-known that comparable objects (previously in the context of the branch, currently more in the context of factors such as sales growth, etc.) are adapted by means of multipliers representing factors such as profit, cash flow, or number of customers [25, 26]. Similar to the local-global principle of the similarity assessment, the adaptation multipliers used in our approach are decomposed into *local* and *global multipliers* (Eqn. 4).

$$c'_p := c_p \cdot \prod_{a \in A} \underbrace{\left(\frac{i_a(q_a)}{i_a(c_a)}\right)^{w_a}}_{local\ multiplier} \quad (4)$$

The global multiplier is computed by the weighted product of the individual local multipliers determined for selected attributes A relevant for adaptation. For each local multiplier, an attribute-specific index function i_a computes a valuation index based on the value of a single attribute $a \in A$. The ratio of the index values for the query attribute q_a and the case attribute c_a determines the local multiplier, additionally taking the attribute weight w_a into account. To adapt the sales price c_p of a case c (c'_p is the adapted price), we consider the attributes "time", "length of the domain name", "number of searches at Google", "occurrence of hyphens", as well as "occurrence of special characters".

Next, an example local multiplier for the attribute "time" is presented. It is based on the index values from the Internet Domain Name Index (IDNX) [1]. This is a price index ranging from 2006 until the current day, which indicates the development of the price level in the domain market. Methodically, neither the average price (vulnerable to outliers) nor the median (does not consider the quality of the traded domain names) is used, but instead the hedonic repeat-sales method [27] is applied. The very rare occurrence of a repeated sale of the same domain serves as the control variable. Otherwise, the method is based on the comparison of very similar domain transactions over time. These are domain transactions that only differ in the TLD, prior to compensating TLD-specific difference in the price [1]. A detailed description of the local multipliers for the other attributes is given by Dieterle [19].

After each of the k-most similar cases C_q is adapted, cases that are considered outliers O_q are detected. Outliers are defined as the cases whose adapted price c'_p deviates "too much" from the median. Here, "too much" means less than half or more than double as high as the median of the k-most similar cases. Outlier cases are not considered in the subsequent computation of the overall appraisal value q_p. It is determined according to (5) by a weighted average of the adapted prices c'_p of the most similar cases using the similarity values of the cases as weights w_c^q.

$$q_p := \frac{\sum_{c \in C_q - O_q} (c'_p \cdot w_c^q)}{\sum_{c \in C_q - O_q} w_c^q} \tag{5}$$

3.5 A Simple Example

The appraisal calculation for the target domain `spanienurlaub.de` (German for "holiday in Spain") in 2012 is shown as an example in Fig. 3. The eleven most similar domain transactions are illustrated on the left side as is the respective year and price. Each case contains one of the words "Urlaub", "Reise" or "Reisen" ("vacation", "trip" or "trips" in German), i.e., a word that belongs to the tourism branch. In each case, this word is combined with a second word, hence the length of the domain name approximately corresponds to the length of the query. Furthermore, most of the cases have a transaction year similar to the query. Six of the eleven cases also include a region name, from which three regions belong to the Mediterranean area. Hence, the similarity measure is able to identify cases from closely related web pages. The right side of Fig. 3 shows the weights, the global multipliers determined during adaptation, as well as the new prices resulting from the adaptation. The results are sorted by price in ascending order and the cases that deviate too much from the median are marked as outliers. They are not considered for the overall price calculation. In this example, the appraised value for the domain `spanienurlaub.de` is 6,060 €.

3.6 Tweaking Support for Professional Users

Aside from the automated appraisal that was just described, the developed approach is also capable of supporting experts in the appraisal of domain names by providing additional means of tweaking the results. This is particularly useful for commercial

domain assessors and publicly appointed and sworn experts with strong knowledge in the field. Using a revision function, the expert user has three options to tweak the results: He/She can change the weighting of the most similar cases manually. If it does not correspond to his/her estimation, he/she can manually change the multipliers when he/she has reasonable grounds to believe that a different value relationship exists between the case and the query. Moreover, he/she can change whether certain cases are considered as outliers or not. The new appraisal value is automatically re-calculated on this basis.

4 Experimental Evaluation

The presented approach is fully implemented as a prototypical system, called Internet Domain Name Appraisal Tool (IDNAT). The implementation was done in JAVA using the object-oriented CBR framework jCOLIBRI2 [28]. In order to perform an experimental evaluation, a case base consisting of 4,231 cases describing domain sales transactions with the TLD .de was extracted from the Internet. For this purpose, the domain transaction list of the United-Domains AG [9] was used, since it is easily accessible and relatively comprehensive with over 1,000 .de entries. Furthermore, the world's largest public list from Namebio [8], with over 3,000 relevant entries, was also used.

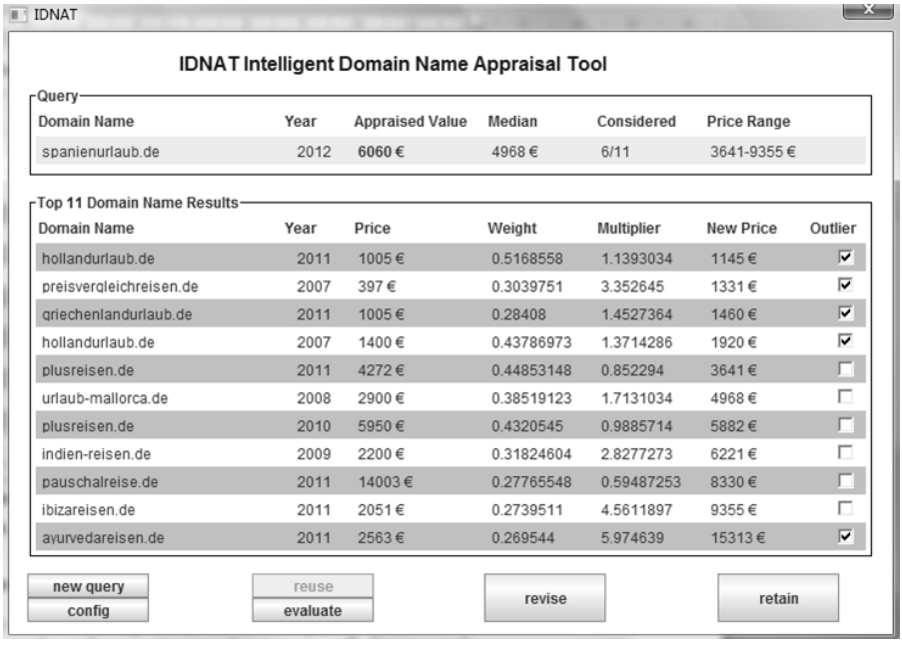

Fig. 3. Price calculation for a query

The purpose of the experimental evaluation was to assess the appraisal quality, i.e., the correlation between proposed appraisal value by IDNAT and the real sales value of previous domain sales transactions. Additionally, we aim to evaluate the impact of the weights used during the similarity assessment. For this purpose, three weighting profiles were defined as shown in Table 2, which differ in the balance between semantic and numeric aspects.

Table 2. Weighting profiles for the evaluation

Experimental design	Basic assumption	Less semantic	More semantic
Semantic aspects	37.04 %	22.73 %	54.05 %
Numerical aspects	62.96 %	77.27 %	45.95 %

The semantic aspects refer to the meaning of the domain name and thus, contain the three attributes “words”, “word components” and “category”. The numerical aspects refer to the number of searches, the number of words, the length, etc., i.e., to all further attributes. The weighting profiles were selected for the evaluation because we observed that, quite often, only a small number of retrieved cases are highly similar with respect to both categories of features.

The evaluation was performed using a leave-one-out test. Hence, each case of the case base is used as query to the system – the case used as query is temporarily removed from the case base. Then, the percentage of correctly appraised queries is determined. An appraisal is defined to be correct if the value is not less than half and not more than double the real sales price. Such a wide range is necessary due to the large price differences that even occur for repeated domain transactions. Furthermore, the correlation coefficient (also called Pearson product-moment correlation coefficient) and the Spearman's rank correlation coefficient are calculated. The evaluation shows how close the system’s appraisal is to the real value and vice versa. Thus, it may also be possible to gain insight as to how fair the actually achieved market prices are. The results are illustrated in Table 3 and Fig. 4.

Table 3. Evaluation Results

Experimental design	Basic assumption	Less semantic	More semantic
Correctly valued	46.02 %	46.84 %	42.68 %
Pearson’s correlation coefficient	0.6729	0.5344	0.6452
Rank correlation coefficient	0.5742	0.5952	0.5214

The evaluation showed that using the profile basic assumption, 46 % of the 4,231 cases were correctly appraised. The average retrieval and adaptation time was 6.1 seconds, which is quite acceptable. The correctly appraised cases are also plotted in the inner interval of Fig. 4. Only 7% of the cases lie outside the outer interval, which are extreme outliers whose estimated value is less than one-tenth or more than ten times the achieved price. For these cases, we are speculating that most of them reflect incorrect sets of data or they are not pure domain transactions, but rather transactions involving entire projects or companies.

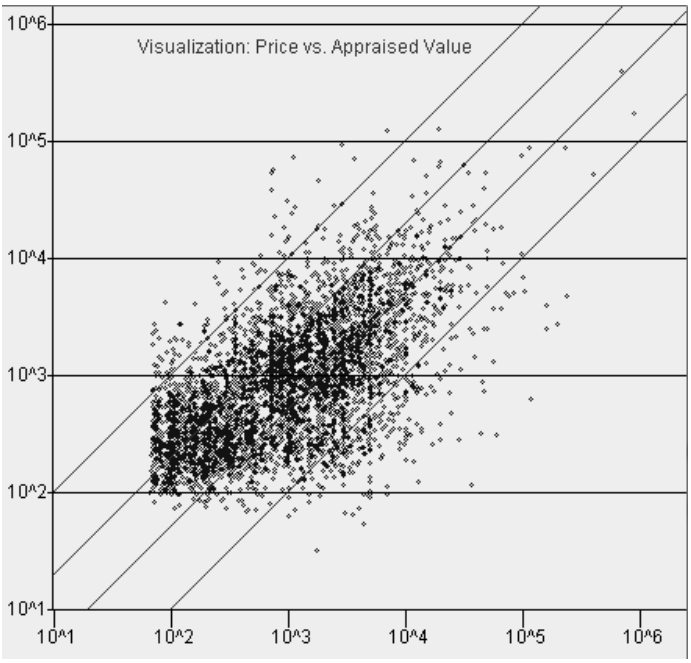

Fig. 4. Visualization of the real price vs. the estimated value (for the basic assumption profile)

The results also demonstrate that the real price and the estimated value have a clear positive correlation. The "More semantic" profile has a negative impact on the quality of the results. This relates both to the defined correctness criteria as well as to the correlation according to both correlation coefficients. On the other hand, the "Less semantic" profile does not deliver clear results. We believe that the reason for the negative impact of the more semantics profile is due to the fact that the prices of Internet addresses within a branch can be very different. If the numeric attributes differ too much and hence, very high or very low adaptation multipliers are used, the adaptation becomes less reliable.

While our evaluation was able to assess our approach in isolation, it does not allow for an assessment of it with respect to currently available assessment methods that follow an idea similar to the market-oriented domain appraisal approach used in Estibot. A comparison with existing approaches has two fundamental problems. First, most available services only allow for a very limited number of free queries per day. Secondly, a leave-one-out evaluation is not possible, since there is no means of explicitly excluding cases during the appraisal. Only a complete re-implementation could solve this problem, which however is difficult, since not all of the details of the appraisal algorithms are disclosed.

5 Conclusion and Future Work

This paper clearly demonstrates that domain names can be appraised using a CBR system. However, the evaluation also unveils the fact that a database of approximately

4,200 cases may be too small to detect a sufficient number of comparable transactions with adequate similarity for a complete TLD. Hence, an important starting point for future work is an enlargement of the case base. This involves including additional domain transaction lists as well as TLDs other than .de. In particular, special considerations are required to enable the adaptation across different TLDs. This is particularly dependent on the language and the popularity of a term in the regions covered by the TLDs. With regards to the similarity between categories, existing links between the same categories in different languages must be considered within the taxonomy.

The Internet administration ICANN intends to allow unlimited TLDs in the near future. It is expected that the amount will be in the four-digit range. It would be interesting not only to include a limited number of TLDs in the model, but to generalize it towards a completely generic approach, accepting arbitrary combinations of SLD and TLD.

The current computation time of a query increases linearly with the size of the case base because a simple sequential retrieval algorithm is used. With this approach, we estimate that the retrieval time would be almost five minutes for a case base of a realistic size of 200,000 transactions. Hence, more efficient retrieval methods are required, e.g., a two-step MACFAC approach.

References

1. Lindenthal, T.: Valuable Words: Pricing Internet Domain Names. Working Paper (2011), http://www.idnx.com/working_paper_IDNX.pdf
2. Lusht, K.M.: Real Estate Valuation: Principles & Applications. IRWIN, Chicago (1997)
3. Cortesi, G.R.: Mastering Real Estate Principles. Dearborn, Chicago (1996)
4. Mettling, S., Cusic, D.: Principles of Real Estate Practice. Performance Publishing Company, Bradenton (1996)
5. Gonzalez, A.J., Laureano-Ortiz, R.: A Case-Based Reasoning Approach to Real Estate Property Appraisal. Expert Systems With Applications 4, 229–246 (1992)
6. Bonissone, P.P., Cheetham, W.: Financial Applications of Fuzzy Case-Based Reasoning to Residential Property Valuation. In: Proceedings of the Sixth IEEE International Conference on Fuzzy Systems, pp. 37–44. IEEE (1997)
7. McSherry, D.: An Adaptation Heuristic for Case-Based Estimation. In: Smyth, B., Cunningham, P. (eds.) EWCBR 1998. LNCS (LNAI), vol. 1488, pp. 184--195. Springer, Heidelberg (1998)
8. Namebio, http://www.namebio.com
9. united-domains AG, http://www.domain-spiegel.de
10. IAC Search & Media, Inc., http://www.teoma.com
11. AOL Inc., http://www.dmoz.org
12. Huber, F., Dingeldey, D.: Ratgeber Domain-Namen, united-domains AG, Starnberg (2001)
13. Benken, B.: http://www.bewertungsformel.de
14. Sunyaev, A., Kaletsch, A., Krcmar, H.: Erfassung, Analyse und Bewertung von Domainnamen: Modellansatz und prototypische Umsetzung. Informatik-Spektrum 35(1) (2012)
15. estibot.com inc., http://www.estibot.com
16. Renshaw, E.F.: Scientific Appraisal. National Tax Journal 11, 314–322 (1958)
17. Do, A.Q., Grudnitiski, G.: A Neural Network Approach to Residential Property Appraisal. The Real Estate Appraiser 58, 38–45 (1992)

18. Czernowski, R.M.J.: Expert Systems in Real Estate Valuation. Journal of Property Valuation and Investment 8(4), 376–393 (1990)
19. Dieterle, S.: Marktorientierte Bewertung von Internet-Domainnamen – Konzeption eines fallbasierten Systems. Bachelor Thesis, University of Trier (2011)
20. Chirita, P.A., Nejdl, W., Paiu, R., Kohlschütter, C.: Using ODP Metadata to Personalize Search. In: SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 178–185. ACM, New York (2005)
21. Kolb, P.: Experiments on the difference between semantic similarity and relatedness. In: Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009, vol. 4, pp. 81–88. Northern European Association for Language Technology (2009)
22. Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics 32, 13–47 (2006)
23. Bergmann, R.:Experience Management: Foundations, Development Methodology, and Internet-Based Applications. LNCS (LNAI), vol. 2432. Springer, Heidelberg (2002)
24. Bunke, H., Messmer, B.T.: Similarity Measures for Structured Representations. In: Wess, S., Richter, M., Althoff, K.-D. (eds.) EWCBR 1993. LNCS, vol. 837, pp. 106–118. Springer, Heidelberg (1994)
25. Boatsman, J.R., Baskin, E.F.: Asset Valuation with Incomplete Markets. The Accounting Review 56, 38–53 (1981)
26. Alford, A.: The Effect of the Set of Comparable Firms on the Accuracy of the Price-Earnings Valuation Method. Journal of Accounting Research 30, 94–108 (1992)
27. Shiller, R.J.: Measuring Asset Values for Cash Settlement in Derivative Markets: Hedonic Repeated Measures Indices and Perpetual Futures. The Journal of Finance 48(3), 911–931 (1993)
28. GAIA – Group of Artificial Intelligence Applications, `http://gaia.fdi.ucm.es`