G-Form: A New Approach for Visual Interpretation of Deep Web Form as Galaxy of Concepts

Radhouane Boughammoura, Lobna Hlaoua and Mohamed Nazih Omri

Faculty of Sciences of Monastir, University of Monastir
Research Unit MARS, Monastir, Tunisia

E-mails: Radhouane.Boughammoura@gmail.com, Lobna1511@yahoo.fr, MohamedNazih.Omri@fsm.rnu.tn

***Abstract*—Deep Web is growing rapidly with multitude of devices and rendering capabilities. Despite the richness of Deep Web forms, their rendering methodology is very poor in terms of capacity of expression. Hence, the user has no indication about the richness of the query and the query capability when he interprets this interface. In this paper, we propose a new rendering approach of Deep Web forms which is easy to interpret by the user and reflects the exact meaning of the query. We have evaluated our algorithm on standard dataset and compared it to a well known state of the art algorithm. Our approach showed good performance with respect to standard measures.**

Keywords- Web Applications; Deep Web; Information Retrieval; Query Interface; Query Interpretation; Galaxy of Concepts; Pertinence of a Concept; Human Computer Interface; Visualization.

I. INTRODUCTION

Deep Web is the part of the Web which is not reachable via hyperlinks [2][4][7][12][14][15]. It is hidden behind Web forms which give access to Deep Web databases. Information on databases is a really big treasure. More than 90% of the information from the Web comes from Deep Web. In addition, this information is very rich in terms of quality of service offered to internet users [12][13][14]. We aim by our job to reveal the Deep Web to novice internet users via new, simple, and easy-to-use Web forms.

A Web form is an information retrieval interface which give access to Deep Web data. It is a graphical representation of the query using a set of fields. Users form their query by visual interpretation of the meaning of the query interface. The design method of the Web form [1][3][5][6][10][11] is very important since it is the only source of inspiration for novice users in order to understand the meaning of the query. A bad interpretation leads to an incorrect query and hence restricts access to Deep Web services. In this paper, we focus on the design aspect of Web forms in order to offer to novice users easy- to-use forms.

In the Web form presented in Fig. 1.(a), we notice presence of white fields crossing the Web form horizontally. These fields are very important; they indicate the presence of semantic entities (or semantic concepts).A novice user may not pay attention to these fields.

We present below the relevance of the white fields.

Let us consider an example in which the user is not interested in non-stop flights but is searching for a flight with one stop to reach the destination. Suppose the user is searching for all flights having as destination ‘’Tunis’’ offered by an airline company with one stop city. The user may give by mistake the destination city and the number of passengers and leave the departure city empty. This request will be wrong unless the user knows all stop cities for destination ‘’Tunis’’. This is not evident and will be a burden for a novice user as he must formulate as many queries as there are stop cities.

In this paper, we present a new design methodology which simplifies the design of the Deep Web form. While our methodology preserves the query capability of the Web form, it removes the complexity of the query with an easy-to-use form containing all necessary and pertinent fields. The resulting form becomes very simple and, more importantly, semantically very rich.

The rest of the paper is organized as follows. Section II presents a brief review of related works. Section III explains the motivation of our new approach G-form. In Section IV, we detail the principle of the G-Form and our experiments will be presented in Section V. Section VI concludes the paper.

II. RELATED WORK

According to the literature, Z-form [9] is considered as the most used form in the Deep Web. Z-Form is a flat query where all fields are listed at the same level of granularity. The name Z-Form comes from the fact that the user reads Z-Form like reading lines in a paragraph: he begins by the first line, then the second, etc. This reading strategy resembles to the letter Z (see Figure 1.b).

 ISBN: 978-1-61208-437-4

Figure 1. Z geometric pattern in a Z-Form

Z-Forms are the most used information retrieval query interface on the Web. Despite their reputation, Z-Forms have drawbacks. The first drawback is that all fields are rendered in the same interface. When the number of fields is large, the query becomes very complex and novice users find it difficult to formulate a correct query. We have seen that a line segment covers an entire line and forms hence one semantic entity of fields. However in many cases, more than one semantic entity may appear in one line. The Z-geometric pattern does not detect all semantic entities in the line, but considers the entire line as one semantic entity.

Ko-Chiu [20] proposes a new approach for effective surfing in the visualized interface of a digital library. This interface is designed for novice users (children).They found that information retrieval seeking of novice users is influenced by their curiosity and hindrance. Ko-Chiu studies the interactions and the usability of various search interfaces, and the enjoyment or uncertainty experienced by children when using virtual game-like interface. When novice users search for information, they have specific directions but they do not have a specific search target. These novice users have no special training or beliefs regarding search strategies. A visual interface based on the navigation experience is used to help users build mind maps. This interface stimulates curiosity of novice users in order to enhance the information retrieval experience.

III. AFFINITIES BETWEEN DEEP WEB AND GALAXY

G-Form is a Deep Web form which is inspired by the concept of a “galaxy. First, we will present the affinities between a galaxy and a Deep Web form, then we explain the way galaxy-form is build.

TABLE 1. ANALOGY BETWEEN DEEP WEB FORM AND GALAXY

Deep Web Form	Cosmic Universe
Field	star
Group of fields	galaxy
Super-group of fields	Super-galaxy
degree of pertinence of field	Distance separating stars
Mean average pertinence of group of fields	Center of mass of galaxy

Hypothesis :

Novice users regard deep web form like they regard cosmic universe.

Stars are fields and a galaxy is one semantic group of fields. When novice users consider the Deep Web form, they move from one semantic entity to another entity just as an astronaut moves from one planet to another or from one galaxy to another in the cosmic universe. If distance is a measure of this cosmic travel, pertinence is the measure of relevance between fields. The center of mass of one galaxy is equivalent to the mean average pertinence of one semantic entity. Sub-sections A and B detail the similarity between a Deep Web form and the galaxy.

A. Web forms

Deep Web forms organize fields respecting a hierarchical schema (see Figure 2). This schema gives the query its meaning. In our previous work [18] we have presented an algorithm, called VIQI (Visual Interpretation of Deep Web Query Interfaces), which is able to extract the hierarchical schema from Z-Form. Figure 2 gives the resulting output of our algorithm when applied to the query interface (left).

The hierarchical schema (see Figure 2) detects the presence of 4 groups: Departure={Leaving From, Going To}, Number of Passengers={Adults, Children, Infants}, Departure Date ={Day of Departure, Month of Departure}, and Returning Date={Returning Day, Returning Month}, and 10 fields: Leaving From, Going To, Adults, Children, Infants, Day of Departure, Month of Departure, Returning Day, Returning Month, and Flight Class) and one super group :root.

As we have mentioned before, the hierarchical aspect of the Deep Web form indicates the presence of semantic relations between entities which are “is-a” and “part-of”’. For this reason, we measure the pertinence of one galaxy of fields as the center of mass of the galaxy.

 ISBN: 978-1-61208-437-4

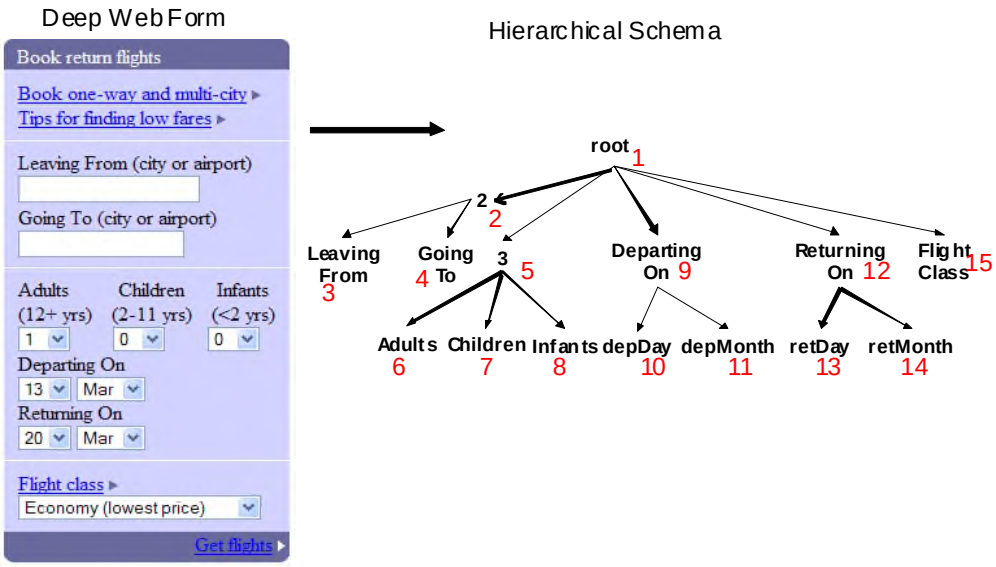

Figure 2. Deep Web form and its hierarchical scheme

Another important aspect in Deep Web forms is degree of pertinence of fields (see red numbers below schema elements of Figure 2). In G-Form, fields are organized according to their pertinence. The most relevant fields are rendered before fields with less pertinence: the most relevant fields are placed on the top left of the Web form, while fields with less pertinence are placed right most on the bottom of the Web form. Hence the field "Adults" is more pertinent than the field "Infants" (pertinence decreases from left to right) and also more pertinent than the field "Children". And field "Going From" is more pertinent than "Flight Class" (pertinence decreases from top to down) as by default users choose economic class while they must mention where they are going in order to formulate a correct query.

The degree of pertinence in G-Form corresponds exactly to the depth first search (DFS) traversal of schema elements. DFS identifies the order of visiting of the schema elements. In Figure 3, we have indicated under each schema element its degree of pertinence. We remark that I < J if element I is rendered before element J in the Web form (on the left). And *DegreeOfPertinence*(I) > *DegreeOfPertinence*(J).

This way, we can measure the relative relevance degree of a group of fields forming one semantic entity. Suppose D is the mean average pertinence degree of a group of fields and r is the distance between the relevance degree of a field and the degree of pertinence of fields group:

if D/ r >ε, then field is relevant in the group

if D/r < ε then field is not relevant

B. Galaxy

Planets in universe are structured according to hierarchical schema like G-Forms. Planets belong to galaxies and there is also super-galaxies grouping a set of galaxies. When we regard the sky by night we observe thousands of brilliant stars (see Figure 3). In reality, each brilliant point is not only a star, but may be another galaxy. As the distance between the observer and galaxy is very large light coming from the galaxy appears like a single point.

We present an example. Let us consider Andaman galaxy, situated at distance r from Earth and having dimension D (see Figure 3) as one point wherever in space as Earth is far away from Andaman galaxy. Andaman galaxy is observed as a single point situated in the center of mass of the galaxy, and has as mass the total mass of the entire galaxy.

In mathematics, quotient D/r is:

$$D/r = \frac{\text{Size of square containing Andaman}}{\text{Distance of Center of mass relative to Earth}} \quad (1)$$

If quotient D/r is very small, we can replace the sum of all stars of galaxy Andaman by only one term situated in the center of mass.

if D/r > ε then star is observed from the galaxy

if D/r <ε then star is observed as a single point

Figure 4 shows planet Earth (on the left) and Andaman galaxy (on the right). The square on the bottom shows a zoom on Andaman galaxy. First, it is clear that, according to an observer inside Andaman galaxy, our galaxy Milky Way may be approximated by a mass point situated at the center of mass. In the galaxy Andaman (or Milky Way) itself, this geometric picture repeats, as indicated in Figure 4. While the quotient D1/r1 is very small, stars situated at the smallest box can be replaced by their center of mass.

 ISBN: 978-1-61208-437-4

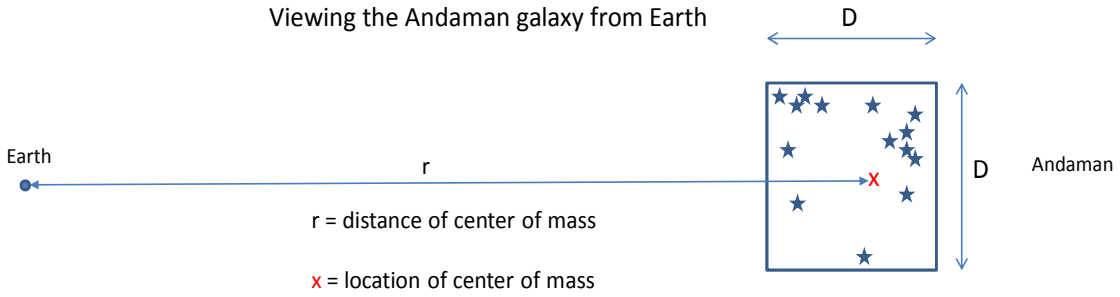

Figure 3. Regarding Andaman galaxy from Earth

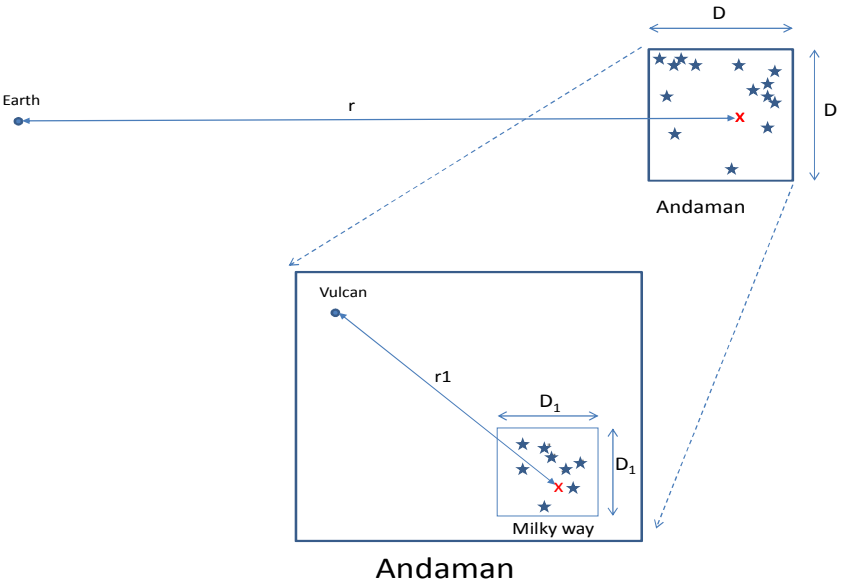

Figure 4. Andaman galaxy analogy

IV. PROPOSED APPROACH: G-FORM

We propose a new approach for visual interpretation of Deep Web forms. The interpretation of the form is based on the hierarchical schema (see Section 3.A) inspired from galaxy. We choose to use a hierarchical representation of the query instead of a flat representation as this representation is richer according to the semantics of the query. The principle of our approach is explained in the following algorithm.

The first Web form is rendered as Z-Form of stars where all fields are stars. When the user clicks on a star, we consider that the user moves to the galaxy containing the field. Our algorithm determines the immediate pertinent fields in the galaxy. A field is considered as pertinent if it is not far from the center of mass (clicked field) of the galaxy: its degree of pertinence is under ε distance, fixed by the user, from the clicked field.

In Table 2, we show which pertinent fields are rendered when the user clicks on a star in Figure 6.(a). For example, let us take ε fixed to 1:

- If the user clicks the field with degree of pertinence equal to 3, then only fields “Leaving From” (with pertinence degree 3) and “Going To” (with pertinence degree 4) are rendered, as they are situated under distance inferior or

 ISBN: 978-1-61208-437-4

```
1)  Procedure  Render_G-Form( {f1, f2, …, fn}, pertinence
    set, ε)
2)  begin
3)     for I from 1 to n
4)        fi ← Not pertinent
5)     enf For
6)     if  clickOn( fj)  then     /* fj is the field clicked */
7)        fj ← pertinent
8)     else
9)        for I from 1 to n
10)          D ←  distanceOfPertinence( fi, fj)
11)          if (D < ε) then
12)             fi ← pertinent
13)          else
14)             fi ← Not pertinent
15)          end If
16)       end For
17)    end IF
18) end.
```

Algorithm 1. Rendering algorithm of G-Form

TABLE 2. OBSERVED FIELDS RELATIVE TO THE CLICKED FIELD

Pertinent fields / user clicks	3	4	6	7	8	10	11	13	14	15
3	*	*								
4	*	*								
6			*	*						
7			*	*	*					
8				*	*					
10						*	*			
11						*	*			
13								*	*	
14								*	*	*
15									*	*

 ISBN: 978-1-61208-437-4

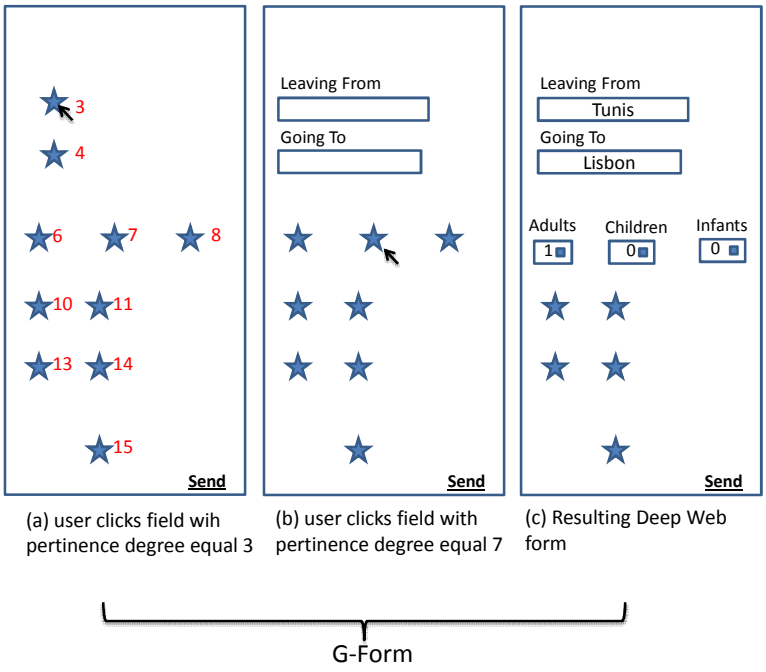

Figure 5. Rendering strategy of G-Form

equal to 1 from the field "Leaving From" (clicked field) having degree of pertinence equal to 3 (see Figure 5.b).

- If the user clicks a field with pertinence 7, then only fields "Adults", "Children", and "Infants" are rendered as they are situated under distance 1 from field "Children" (clicked field) having degree of pertinence equal to 7 (see Figure 5.c).

In Table 2, the rows correspond to the clicked star and columns correspond to the rendered fields. According to Table 2, we remark that the rendered fields depend on the clicked field because, when the user clicks a star, we consider that the observer moves to the galaxy of this star. Hence, only stars at quotient $D/r > \varepsilon$ (see Section 4.B) are observed from this galaxy and all the others are observed as brilliant stars.

V. Experimental Results

Our approach renders the query according to its semantic representation (schema of the query). We have tested the performance of G-Form on a standard dataset ICQ [19]. ICQ is a collection of query interfaces collected from the Deep Web services. For each query interface, its manually extracted query schema is available on dataset. Interfaces are collected into five classes of subjects: Airfare, Automobile, Books, Real estate, and Jobs.

Our evaluation methodology is as follows. We build G-Form for every schema of Web form available on the dataset. Then, we build Table 2 which simulates the user clicks on different stars in the Web form. Then, we count the number of correct entities (group, super-group of fields). An entity is considered as correct if it is semantically coherent. For example entity {Adults, Children, Infants} is a correct entity as it describes the number of passengers. While entity {Time, Adults, Children} is not correct because Date of flight and Number of passengers overlap.

We count for each G-Form number of extracted entities, number of extracted entities which are correct, and then we measure precision, recall, and F1 of the algorithm.

$$Recall = \frac{number\ of\ extracted\ entites}{total\ number\ of\ entites\ in\ the\ Web\ form} \quad (2)$$

$$Precision = \frac{number\ of\ extracted\ \ pertinent\ entities}{\text{number of extracted entities}} \quad (3)$$

$$F1 = \frac{2\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

The experimental results are summarized in Table 3 with ε equals 1:

TABLE 3. EXPERIMENTAL RESULTS

Domain	Airfare	Auto	Books
#extracted	214	102	108
#extracted &pertinent	146	78	70
#total_ entities	200	105	110
Precision	0,68	0,76	0,64
Recall	0,73	0,74	0,63

 ISBN: 978-1-61208-437-4

Figures 6, 7, and 8 show that our approach performs with better results on the domain of interest ‘’Auto’’. The ‘’Auto’’ domain contains flat queries, i.e fields are organized at the same level. Choosing a good ε coefficient makes the visual representation of the query very easy.

Our approach allows to achieve 73% of recall for ‘’Airfare’’ domain. Queries in this domain are hierarchical with many levels. Choosing a small ε coefficient renders the concepts in the “Airfare” domain as small galaxy composed of 2 or 3 fields.

TABLE 4. COMPARISON BETWEEN OUR APPROACH (G-FORM) AND Z-FORM

	Domain	Precision	Recall	F1
Our approach	Airfare	**0,68**	**0,73**	**0,70**
	Auto	**0,76**	**0,74**	**0,75**
	Book	**0,64**	**0,63**	**0,64**
Z-Form	Airfare	0,66	0,70	0,67
	Auto	0,72	0,71	0,71
	Book	0,62	0,60	0,60

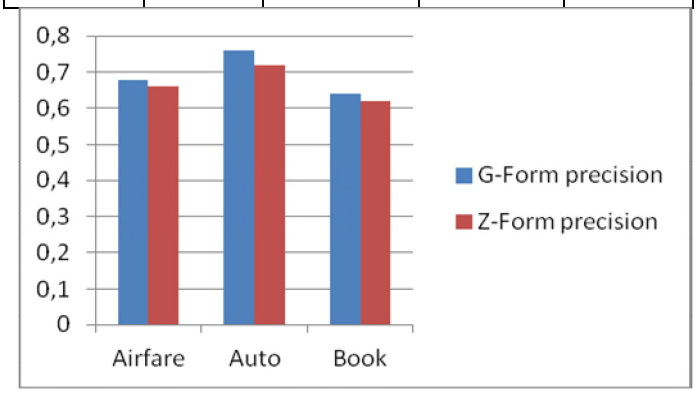

Figure 6. Comparison of precision of our approach and Z-Form

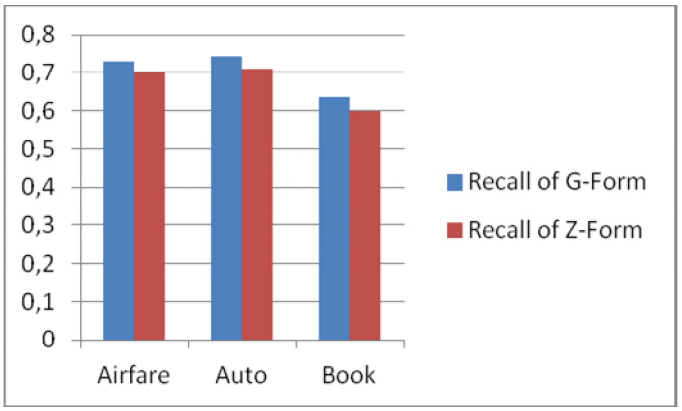

Figure 7. Comparison of recall of our approach and Z-Form

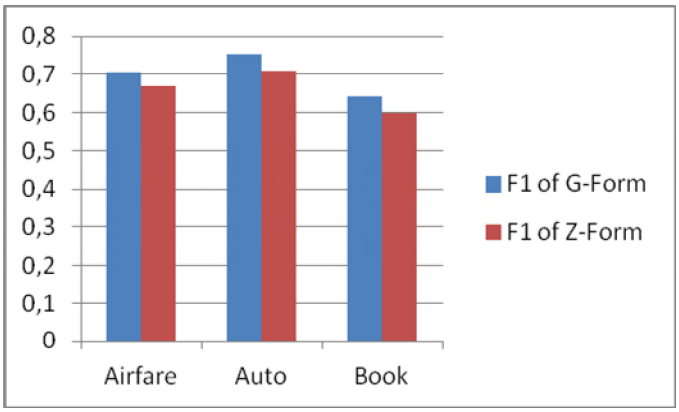

Figure 8. Comparison of F1 of our approach and Z-Form

Figures 6, 7 and 8 summarize the results shown in Table 4. We notice that, with regard to all measures, the two curves corresponding to the two approaches follow the same pace. This phenomenon can be explained by the fact that the process of rendering of fields is based on the query schema which is common to the two approaches. However, Figures 6, 7 and 8 show that the performance of our approach is always superior to the performance of Z-Form. Our approach attends its maximum precision for ‘’Airfare’’ domain as queries in this domain are hierarchical and well adapted to rendering the strategy of our algorithm.

TABLE 5. PRECISION FOR DIFFERENT QUERY COMPLEXITIES

	Precision		
	ε=1	ε =2	ε =3
Airfare	**0,738**	0,649	0,635
Auto	0,732	0,693	**0,792**
Book	0,637	0,601	**0,712**

The coefficient ε is an indicator of the complexity of the query. For small values of ε, only a small group of fields are rendered; the other groups are rendered as stars. This is the case of the ‘’Airfare’’ domain, which is considered as the most complicated domain. For a large coefficient ε, a large group of fields are rendered and interpretation is close to Z-Form. This is the case of the “Book” domain, which is formed with flat queries. The complexity measure of the query can be shown for epsilon 3: “Airfare” is the most complex, ”Book” is relatively more complex than airfare, and “Auto” is the simplest query.

 ISBN: 978-1-61208-437-4

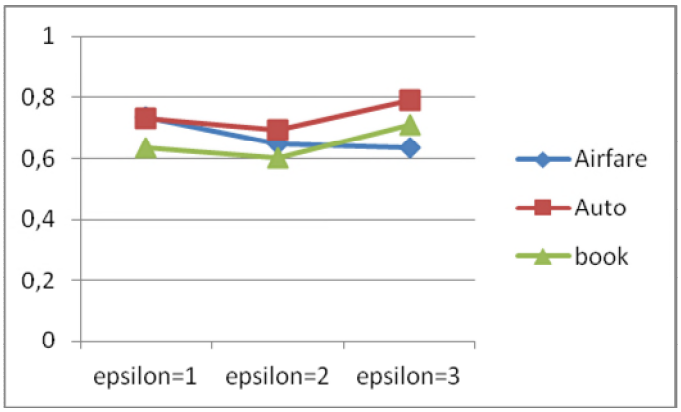

Figure 9. Precision for different query complexities

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a method to solve the complexity of queries in Deep Web forms. We proposed a new design approach G-Form inspired from the concept of a galaxy. It offer to novice users easy-to-use Deep Web forms and reveals clearly the exact meaning of the query even if its schema is complex. There is a strong analogy between the concept of a galaxy and Deep Web form with respect to structure and granularity of entities in each concept.

G-Form is better than Z-Form, which is a well known state of the art algorithm. Z-Form lacks design expressivity because there is no background concerning the semantic value of the query, while our approach is based on a hierarchical schema, which reveals clearly the semantic of the query. Our approach clearly has better performance than Z-Form with respect to precision, recall, and F1 measures of performance.

REFERENCES

[1] G. Agarwal, G. Kabra, K. C-C Chang, “Towards rich query interpretation: walking back and forth for mining query templates”, In proceedings of the international conference on world wide web, 2010.

[2] F. Jiang, L. Jia, W. Meng, X. Meng, ‘’MrCoM: A Cost Model for Range Query Translation in Deep Web Data Integration’’, In Proceedings of SKG '08, 2008.

[3] Z. Zhang, B. He, and K. C-C Chang, ‘’Light-weight domain-based form assistant: querying web databases on the fly’’, In Proceedings of VLDB '05, 2005.

[4] Z. Zhang, B. He, and K. C.-C. Chang , ‘’On-the-fly Constraint Mapping across Web Query Interfaces’’. In Proceedings of VLDB-IIWeb'04, 2004.

[5] J. Jansen and Dick C.A. Bulterman, ‘’Enabling adaptive time-based web applications with SMIL state’’, In Proceedings of DocEng '08, 2008.

[6] M. Jayapandian, H. V. Jagadish. 2008, ‘’Expressive query specification through form customization’’, In Proceedings of EDBT '08, 2008.

[7] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, A.Y. Halevy, ‘’Google’s Deep Web crawl’’, In Proceedings of VLDB, 2008.

[8] E-C. Dragut, T. Kabisch, C. Yu, U. Leser, ‘’A hierarchical approach to model web query interfaces for web source integration’’, In Proceeding of. VLDB 2009, 2009.

[9] W. Wensheng, A-H Doan, C. Yu, W. Meng, ‘’Modeling and Extracting Deep-Web Query Interfaces’’, In Proceedings of AIIS 2009, 2009.

[10] Z. Zhang, B. He, and K. Chang, ‘’Understanding Web query interfaces: Best-effort parsing with hidden syntax’’, In Proceedings of SIGMOD’04, 2004.

[11] Heidinger, C., K. Bohm, Buchmann E., and Spoo M. "Efficient and secure exact-match queries in outsourced databases." World Wide Web: 1-39, 2013.

[12] L.T.H.Vo, J. Cao, and W. Rahayu, ‘Structured content-based query answers for improving information quality’’, In World Wide Web: 1-24, 2014.

[13] J. Losada, J. Raposo, A. Pan, P. Montoto. "Efficient execution of web navigation sequences." World Wide Web: 1-27, 2013.

[14] R. Boughammoura, MN. Omri, and H. Youssef, ‘’Fuzzy Approach for Pertinent Information Extraction from Web Resources’’, Journal of Computing and e-Systems, Vol. 1 No. 1, 2008.

[15] R. Boughammoura, MN. Omri, ‘’Statistical Approach for Information Extraction from Web Pages’’, In proceedings of International Symposium on Distance Education (EAD’2009), 2009.

[16] R. Boughammoura, MN. Omri, ‘’SeMQI: A New Model for Semantic Interpretation of Query Interfaces’’, In Proceedings of NGNS’11, 2011.

[17] R. Boughammoura, MN. Omri, Hlaoua, L. ‘’ VIQI: A New Approach for Visual Interpretation of Deep Web Query Interfaces’’, ICITeS 2012.

[18] R. Boughammoura, MN. Omri, Hlaoua, L. ‘’Information Retrieval from Deep Web based on Visuel Query Interpretation’’, International Journal of Information Retrieval Research, 2(4), 45-59, 2013.

[19] The UIUC Web Integration Repository, Computer Science Department, University of Illinois at Urbana-Champaign, http://metaquerier.cs.uiuc.edu/repository, 2003.

[20] Ko-Chiu Wu, Affective surfing in the visualized interface of a digital library for children, Information Processing & Management, Volume 51, Issue 4, July 2015, Pages 373-390

 ISBN: 978-1-61208-437-4