یکپارچه سازی داده های XMLفازی ناهمگن بر اساس شباهت ساختاری و معنایی

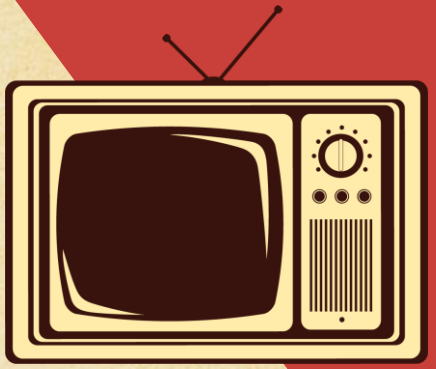
Heterogeneous fuzzy XML data integration based on structural and semantic similarities

Amir Shokri

Dr. RahmaniManesh

نویسندگان

Zongmin Ma
Zhen Zhao
Li Yan



ScienceDirect
www.elsevier.com/locate/fss

اطلاعات

ارسال

23 فوریه 2017

اصلاح و ارسال

6 آپریل 2018

تایید

30 آپریل 2018

1

بخش اول

بخش اول

معرفی XML

XML مجموعه ای از قواعد برای تولید مستندات به فرمی که قابلیت خوانده شدن با ماشین را داشته باشد است این قواعد بر اساس مشخصات استاندارد XML نسخه 1.0 (کنسرسیوم وب)، سایر استانداردهای مرتبط و استانداردهای باز (open source) رایگان تعریف شده اند.

اهداف طراحی XML، بر سادگی، عمومیت و قابلیت استفاده آن در اینترنت تاکید دارند.

XML، قالب داده ای متنی با پشتیبانی قوی Unicode برای زبان های مختلف دنیا است

همچنین طراحی XML بر روی مستندات تمرکز دارد و بصورت گسترده برای ارایه ساختارهای داده ای دلخواه مورد استفاده قرار می گیرد. مانند Web Serviceها.

بخش اول

معرفی XML

بسیاری از APIها به گونه ای نوشته شده اند که برنامه نویسان نرم افزار بتوانند از آنها برای پردازش داده های XML استفاده کنند و Schema های متعددی برای کمک به تعریف زبان های بر پایه XML وجود دارد.

تا کنون صدها زبان بر پایه XML تولید شده است مانند RSS، ATOM،SOAP،XHTML، قالب های برپایه XML، به قالب پیش فرض خیلی از نرم افزارهای اداری تبدیل شده است مانند :

OpenOffice.org(OpenDocument)

Microsoft Office(Office Open XML)

بخش اول

توصیف اولیه XML

- Xml، مخفف eXtensibile Markup Language است.
- Xml برای ذخیره و انتقال اطلاعات است و با HTML فرق دارد و هیچ کدام نمی توانند جایگزین هم باشند.
- Xml به صورت خودتعریف است یعنی نام تگ ها قاعده ی از پیش تعریف شده مانند HTML ندارند.
- Xml در خیلی از جنبه های توسعه وب مانند ساده سازی ذخیره و به اشتراک گذاری اطلاعات استفاده می شود.
- Xml را در سیستم های کامپیوتری به عنوان بانک های اطلاعاتی متنی استفاده می کنند و قابل اشتراک داده ها بین زبان های مختلف برنامه نویسی را دارد.

بخش اول

ساختار یک سند XML

```
<?xml version=”1.0” encoding=”utf-8” ?>
<person id=“1”>
        <firstname>Amir</firstname>
        <lastname>Shokri</lastname>
        <stcode>9811920009</stcode>
</ person >
```

بخش اول

قواعد نحوی XML

- همه المان ها باید تگ بسته داشته باشند.
- اعلان اولیه ی Xml که تعریف نسخه ی سند Xml است نیاز به تگ بسته ندارد.
- Xml به بزرگی و کوچکی حروف حساس است.
- المان های تو در تو باید به هر ترتیبی که باز شده اند بسته شوند.
- مقدار خصیصه های هر برچسب باید بین " " باشد.
- کاراکتر < و & در Xml مجاز نیست.
- Xml چند فاصله ی خالی پشت سر هم را دست نخورده باقی می گذارد.
- قرار دادن کامنت بین متن Xml مانند Html است. مثال :

- <!– Comment →

بخش اول

چکیده

انضمام داده های وب به یک نیاز اساسی برای مدیریت داده های وب تبدیل شده است. تعداد قابل توجهی رویکرد برای یکپارچه سازی داده های زبان تعمیم یافته ی نشانه گذاری XML از منابع اطلاعاتی ناهمگن بدست آمده است.البته این رویکردهای هیجان انگیز هنوز مناسب ادغام اطلاعات XML فازی نیستند به دلیل مشخصات فازی که دارند.در این مقاله،ما یک چهارچوب برای مقابله با ادغام اسناد XML فازی آماده کرده ایم. ابتدا یک مدل درختی جدید برای XML فازی سپس الگوریتمی موثر بر اساس فاصله ویرایش شده ی درخت برای شناسایی ساختار و شباهت های معنایی بین اسناد فازی حاضر در مدل درختی XMLفازی آماده کرده ایم.

بخش اول

چکیده

در مرحله ی بعد یک استراتژدی اتحاد که برای یکپارچه سازی اسناد فازی از منابع داده های مختلف استفاده شده است معرفی کرده ایم و در آخر ما آزمایش هایی برای مشخص نمودن رویکردمان که می تواند به طور موثر اسناد XML فازی را متحد کند انجام داده ایم.

بخش اول

مقدمه

مزیت اصلی XMLباز بودنش بر اساس استانداردها و ویژگی ها می باشد. منعطف بودنش اجازه می دهد اسناد بسط پیدا کنند برای توصیف محتوا در هر سایزی. خیلی از ارگانها و سازمانها از XMLبه عنوان یک توصیف اطلاعاتی و یا مدل مخزن مانند در کاربردهای بر اساس وب خود استفاده می کنند.

انضمام داده های XMLیک امر حیاتی و سختی است به دلیل پیچیدگی و انعطاف پذیری مشخصاتش. داده ی XML انضمام بخشیده می شود با مستقر شدن در منابع اطلاعاتی ناهمگن و توزیع شده. انضمام داده های XMLبه این معنا نیست که مستقیما منابع اطلاعاتی XMLرا در کنار یکدیگر قرار بدهیم . برای موفقیت در یکپارچه سازی اطلاعات XMLنیاز است تا تناقضها و سپس همپوشان ها را از اسناد شناسایی و حذف کنیم.

بخش اول

مقدمه

یک رویکرد برای یکپارچه سازی اطلاعات XMLاین است که یک انبار اطلاعاتی بسازیم که یک ظرف اطلاعاتی از انضمام و یکچارچگی از منابع اطلاعاتی ناهمگن تولید کند.

هدف اصلی آن است که پروسه ی شناسایی و تطبیق تناقض ها و همپوشانی ها ازمقدارهای استخراج شده XMLمحلی را آسان بنماید. از انجا که ساختار اسناد XML ناهمگن است یک تناقض در ساختار می تواند باعث سختی انضمام سازی شود.

رویکردهای فعلی یکپارچه سازی فرض می کنند که اطلاعات XMLاشیاء مشخص در جهان واقعی را مشخص می کنند.

مثال

<name>George</name>

<Dist Type = "disjunctive">

<Val Poss = 1.0>

<position>Associate Professor</position>

</Val>

<Val Poss = 0.7>

<position>Lecturer</position>

</Val>

</Dist>

بخش اول

مقدمه

یک مشکل ریشه ای که نیاز به حل شدن در انضمام اطلاعات XMLفازی ناهمگن دارد، شناسایی همان و یا شباهت های عنصرها و مقادیر است.

براین اساس تناقض ها و همپوشانی های عناصر حذف می شوند. از این رو، پیدا کردن تناقض ها و همپوشانی های عناصر و مقادیرشان در داده های XMLفازی برای یکی کردن چندین اسناد XMLفازی به یک سند، نیاز است.

برای یکپارچه سازی اطلاعات XMLفازی از منابع اطلاعاتی ناهمگن، ما یک رویکرد موثر و مفید برای تقویت یکپارچگی سند XMLفازی آماده کردیم.

بخش اول

مقدمه

اولین قدم را در راستای ساختن مدل جدید درختی XMLفازی برمی داریم. این کار توصیف داده XMLفازی را و تسخیر ساختار و اطلاعات معنایی را نیز ساده می سازد. رویکرد موثری را برای تشخیص شباهت ها بین گره ها درخت XMLفازی آماده می کنیم و بر روی ساختار و مقایسه شباهت های معنایی اسناد ناهمگن XMLفازی که از منبع های مختلف جمع آوری شده اند، تمرکز کرده ایم. پاسخ های اطلاعاتی در گره درخت ها شناسایی می کنیم و در مورد شباهت های گره ها از درخت های XMLفازی مختلف تصمیم می گیریم ارتباط بین دو سند XMLفازی می تواند از گره هایی پاسخ هایشان تشخیص داده شود می توانیم مستقیما تناقض ها و همپوشانی ها بین اطلاعات XMLفازی که از منابع مختلف جمع آوری شده اند مشخص کنیم.

02

بخش دوم

بخش دوم

کار مرتبط

- XMLناقص
- تطبیق موجودیت XML
- انسجام اطلاعات XML

3

بخش سوم

بخش سوم

مدل درختی XMLفازی

- اسناد XMLفازی
- مدل های درختی سند xmlفازی

بخش سوم

```
1.    <Teaching>
2.        <Teacher Tid = “007102”>
3.          <Dist Type = “disjunctive”>
4.            <Val Poss = 1.0>
5.              <Position>Associate Professor</Position>
6.            </Val>
7.            <Val Poss = 0.7>
8.              <Position>Lecturer</Position>
9.            </Val>
10.         </Dist>
11.         <Name>George</Name>
12.       </Teacher>
13.       <Student Sid = “20130111”>
14.         <Dist Type = “conjunctive”>
15.           <Val Poss = 0.6>
16.             <Email>Mary@hotmail.com</Email>
17.           </Val>
18.           <Val Poss = 1.0>
19.             <Email>Mary@gmail.com</Email>
20.           </Val>
21.         </Dist>
22.         <Name>Mary</Name>
23.       </Student>
24.   </Teaching>
```

شکل 1 : اولین قطعه سند فازی

بخش سوم

معرفی FXTM

یک درخت دستوراتی T(N, E) می باشد، N و E دستورهایی از گره ها و مرزهای سند XML فازی T هستند.

در FXTM یک پنج تایی است :

(NodeLabel, NodeDepth, NodeFuzzy, NodeType, NodePoss)

node label : برچسب اسم گره / صفت است.

node depth : عمق تودرتوی عنصر/ ویژگی در سند xml است.

node fuzzy : برای اینکه یک گره فازی است یا کریسپ استفاده می شود اگر مقدار گره فازی 1 باشد گره مربوط به گره فازی است و اگر مقدار آن 0 باشد گره مربوط به گره کریسپ است. ارزش گر فازی برای یک گره ی Dist یا Val صفر است.

node type : بیانگر نوع تفسیری یک گره فازی است که متصلی یا منفصلی است. یک گره کریسپ ارزش پیش فرض گره Dist اش مقدار null است.

Node poss : گره ی Poss درجه عضویت یا ارزش یک گره را می دهد و مقدار پیش فرض برای گره ی کریسپ null می باشد.

```
1.  <Teaching>
2.      <Teacher Tid = “007102”>
3.        <Dist Type = “disjunctive”>
4.          <Val Poss = 1.0>
5.            <Position>Associate Professor</Position>
6.          </Val>
7.          <Val Poss = 0.6>
8.            < Position>Professor</Position>
9.          </Val>
10.       </Dist>
11.       <Name>George</Name>
12.       <Office>B208</Office>
13.     </Teacher>
14.     <Student Sid = “20130425”>
15.       <Dist Type = “conjunctive”>
16.         <Val Poss = 0.7>
17.           <Email>John@hotmail.com</Email>
18.         </Val>
19.         <Val Poss = 1.0>
20.           <Email>John@gmail.com</Email>
21.         </Val>
22.       </Dist>
23.       <Name>John</Name>
24.     </Student>
25. </Teaching>
```

شکل 2 : دومین قطعه سند فازی

بخش سوم

اسناد XML فازی

اسناد Xmlفازی شامل چهار نوع عنصر یا ویژگی جدید :

(Type, Val, Poss, Dist)

می شود که مقادیر آنها شامل اطلاعات فازی می شود.

بخش سوم

تبدیل مقادیر فازی به گره های مربوطه

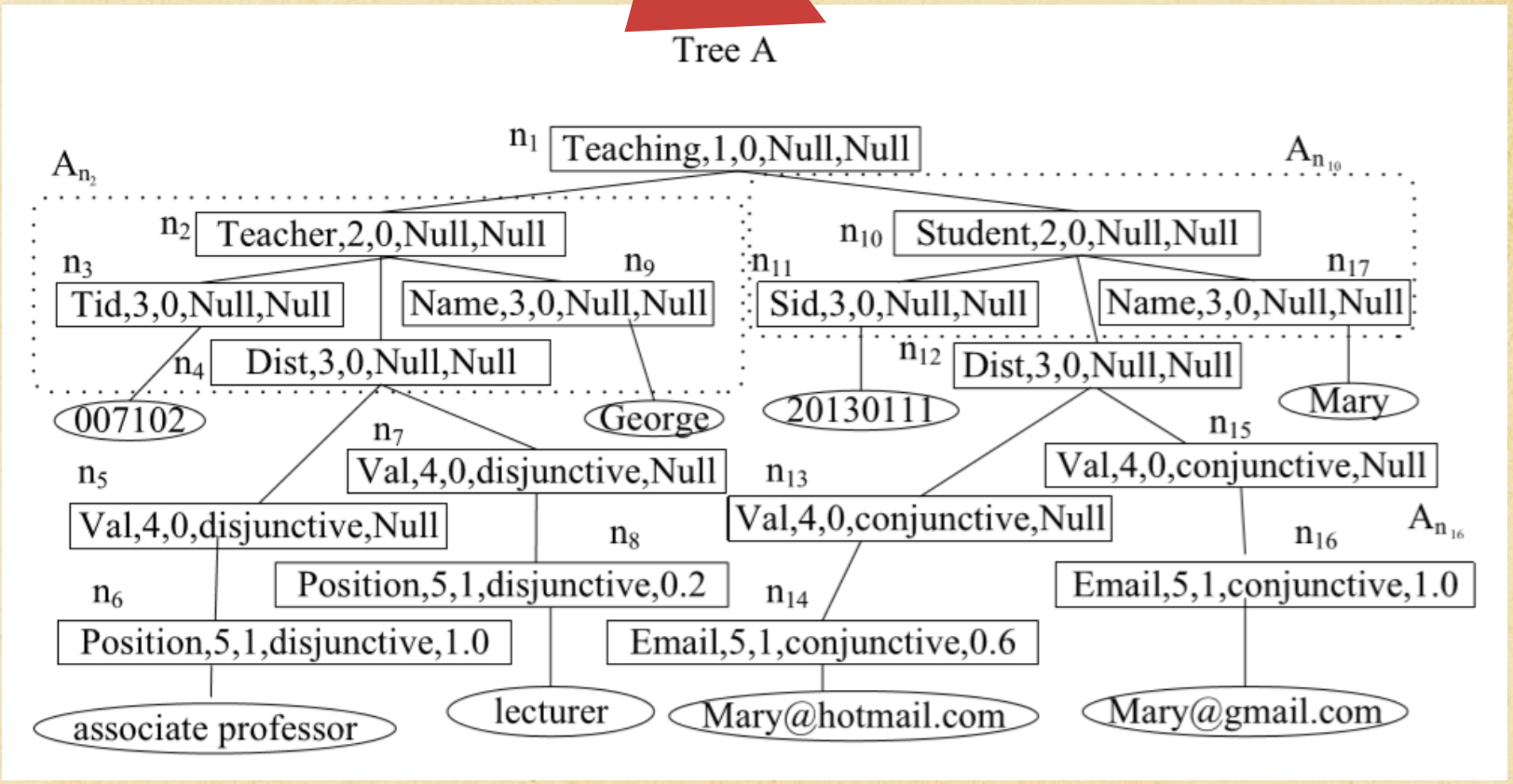
مرحله 1 : ما مقادیر NodeFuzzy و NodeType و NodePoss از والدین آنها به ارث می بریم.

مرحله 2 : برای گره Type ما نیاز به یک کپی از مقدار Type Node و Node Type از خواهر و برادر های Type Node داریم که Type Node را از زیردرخت آن حذف می کنیم.

مرحله 3 : برای گره IVal اگر فقط گره Poss به عنوان گره فرزند باشد ما گره Val و زیردرخت آنرا حذف می کنیم.

مرحله 4 : برای گره Poss ما یه کپی از مقدار گره Poss در NodePoss و خواهر و برادرهای گره NodePoss می دهیم و گره Poss و زیردرخت های آنرا حذف می کنیم.

مرحله 5 : برای گره IDist اگر پس از پردازیش فوق به گره برگ تبدیل می شود و ما گره Dist و زیردرخت های آنرا حذف می کنیم.



شکل 3 : مدل دریخت یک سند XMLفازی مربوط به سند XMLفازی در شکل 1

4

بخش چهارم

بخش چهارم

اقدامات مشابه

- اقدامات شباهت معنایی گره
- اندازه گیری نام برچسب
- اندازه گیری اطلاعات فازی
- اقدامات تشابه ساختاری زیرسطحی
- پیچیدگی های کلی

بخش چهارم

اندازه گیری نام برچسب

$$Sim_{labelSyn}(L_1, L_2) = \frac{1}{1 + editDistance(L1, L2)}$$

$$Sim_{labelSem}(L_1, L_2, S_N) = \frac{2Log\ p(l_0)}{\log p(L_1) + \log p(L_2)}$$

$$Sim_{labelSem}(L_1, L_2) Sim_{label}(L_1, L_2) = Sim_{labelSyn}(L_1, L_2) + (1 - \gamma)$$

بخش چهارم

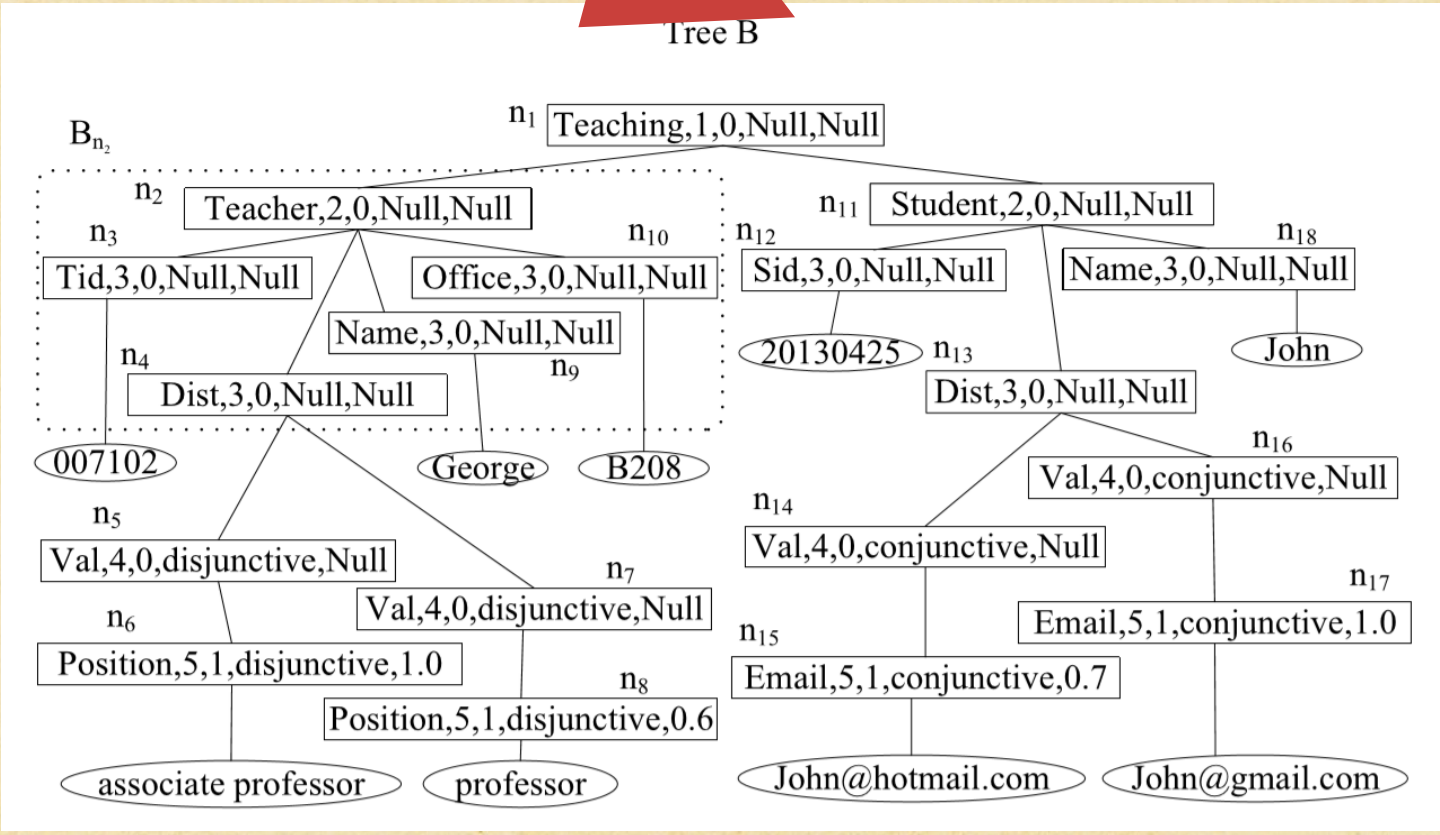
اندازه گیری اطلاعات فازی

$$Sim_{Type}(T_1, T_2) = \begin{cases} 1 & if\ T_1 = T_2 \\ 0 & if\ T_1 \neq T_2 \end{cases}$$

$$Sim_{Poss}(P_1, P_2) = 1 - |P_1 - P_2|$$

$$Sim_{Node}(N_1, N_2) = \omega_L \times Sim_{Label(L_1,L_2)} + \omega_T \times Sim_{Type}(T_1, T_2) + \omega_P \times sim\_Poss(P_1, P_2)$$

were wL, wT , and wP are provided as weight coefficients, and $\omega_L + \omega_T + \omega_P = 1$.



شکل 4 : یک مدل درخت سند XMLفازی مربوط به سند XMLفازی شکل 4

بخش چهارم

اقدامات تشابه ساختاری زیرسطحی

- تعریف شماره 2 (درخت مرتب)
- تعریف 3 (زیر لایه ی دولایه)
- تعریف 4 (زیر شاخه گره برگ)
- تعریف 5 (گره درج)
- تعریف 6 (حذف گره)
- تعریف 7 (ویرایش اسکریپت)
- تعریف 8 (فاصله ویرایش درخت)
- تعریف 9 (گره های همسان)
- تعریف 10 (فاصله ویرایش زیر لایه های دولایه, TSED)

بخش چهارم

Algorithm 1 TSED.

```
Input:  A, B                                  //Two-layer subtrees to be compared
Output: TSED(A, B)                            //Edit distance between A and B
Begin
1    M = Degree(A)                            //The degree of two-layer subtrees of A
2    N = Degree(B)                            //The degree of two-layer subtrees of B
3    Dist[][] = new[0...M][0...N]
4    If (Sim_Node(A[0], B[0]) > θ)            //Node matching
5    {
6      Dist[0][0] = 0;                        //Structurally matching nodes are associated with a cost of 0
7      MatchNodeNumber add 1                  //Counting matching nodes
8    }
9    Else
10   {
11     Dist[0][0] = 1 //Unit costs
12   }
13   For (i = 1 to M) {Dist[i][0] = Dist[i−1][0] + Cost_DelNode(A[i])}      //Cost_DelNode(A[i]) is a cost of 0 or 1
                                               //Total cost of deleting all nodes in the source document tree
14   For (j = 1 to N) {Dist[0][j] = Dist[0][j−1] + Cost_InsNode(B[j])}      //Cost_InsNode(B[j]) is a cost of 0 or 1
                                               //Total cost of inserting all nodes in the destination document tree
15   For (i = 1 to M)
16   {
17     For (j = 1 to N)
18       {                                     //Identifies the set of insertion/deletion operations having the minimum overall cost
19       Dist[i][j] = Min{
20                    {IF (Sim_Node(A[i], B[j]) > θ) {Dist[i−1][j−1], MatchNodeNumber add 1}
21                      else {Dist[i − 1][j − 1] + 1}},
22                    Dist[i−1][j] + Cost_DelNode(A[i])      //Cost_DelNode(A[i]) is a cost of 0 or 1
23                    Dist[i][j−1] + Cost_InsNode(B[j])      //Cost_InsNode(B[j]) is a cost of 0 or 1
24                   }
25       }
26   }
27 Return Dist[M][N]
```

الگوریتم 1 : TESD

بخش چهارم

الگوریتم TESD

$$\alpha = \frac{1}{1 + \frac{MatchNodeNumber}{Max(|A|.|B|)}}$$

$$Sim_{FXST}(A,B) = \frac{|A| + |B| - TESD(A,B) \times \alpha}{|A| + |B|}$$

Sim_{FXST} (A, B) ∈ [0, 1].

بخش چهارم

الگوریتم TESD

Sim_{FXST} (A, B) = 1

Sim_{FXST} (A, B) = 0

Sim_{FXST} (A, B) = Sim_{FXST} (B, A).

TSED(An2 , Bn2) = 1

Dist[0][0] = 0, having R(A_{n2}) and R(B_{n2}) matching.

بخش چهارم

پیچیدگی کلی

$$O(|A| \times |B| \times |SN| \times Depth(SN))$$

$$O(|A|||B|)$$

5

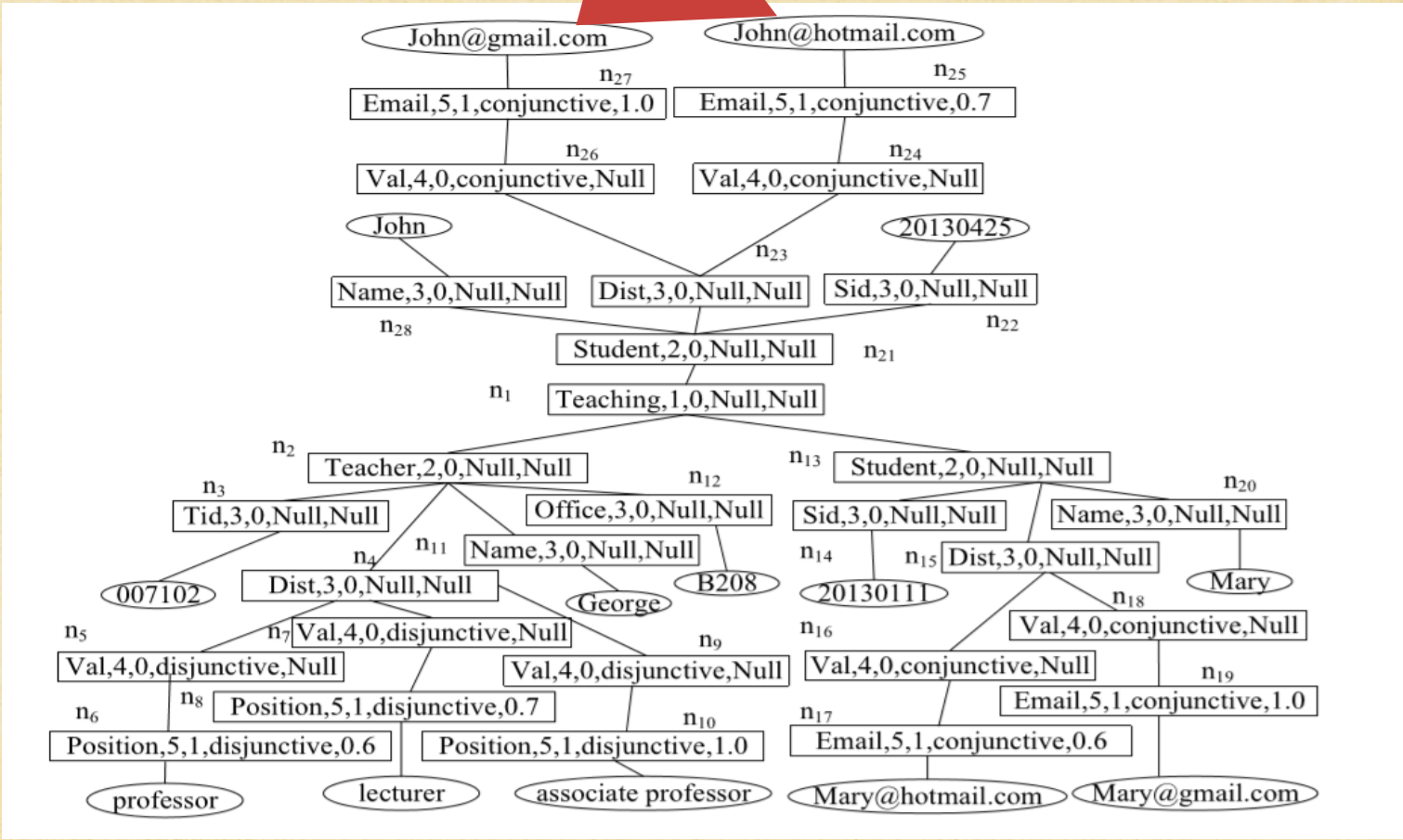
بخش پنجم

بخش پنجم

ادغام داده های XMLفازی

- حل تعارض معنایی و ساختاری
  - تضادهای نامگذاری عنصر/ویژگی
  - تضاد در ارزش ها را مشخص کنید
  - تضادهای ارزش فازی
- ادغام اسناد XMLفازی
- پیچیدگی های کلی

بخش پنجم



شکل 5 : درخت درخت سند XMLفازی ((FXTMدرخت مربوط به سند XMLفازی یکپارچه

بخش پنجم

Algorithm 2 IFXD.

Input: FD_b, FD_f //Fuzzy XML documents FD_b and FD_f that are to be integrated
Output: FD_t //Integrated Fuzzy XML documents FD_t
Begin
$FT_b = \text{FXTM}(FD_b)$ //FXTM FT_b from FD_b
$FT_f = \text{FXTM}(FD_f)$ //FXTM FT_f from FD_f
For (each n_{b_i} in FT_b) //Two-layer subtree rooted in node n_{b_i}
{For (each n_{f_j} in FT_f) //Two-layer subtree rooted in node n_{f_j}
{ do while (!Empty(FT_f))
{If ($Sim_{FXST}(n_{b_i}, n_{f_j}) > \delta$) //Two-layer subtree similarity degree computation
If (Isleafnodesubtree (n_{b_i}, n_{f_j})) //Subtree rooted in n_{b_i}, n_{f_j} is leaf node subtree
If ($Sim_{Node}(n_{b_i} \rightarrow \text{child}, n_{f_j} \rightarrow \text{child}) > \theta$) //Value equality
{$n_{b_i} \rightarrow \text{child.Poss} = \min(n_{b_i} \rightarrow \text{child.Poss}, n_{f_j} \rightarrow \text{child.Poss})$
deltreesubtree(n_{f_j});} //Deleting subtree rooted in n_{f_j} of the foreign tree
Else //Inserting leaf nodes
IFXD($n_{b_i}, n_{f_j} \rightarrow \text{sibling}$) //Integrate sibling subtree of $n_{f_j}, n_{f_j} \rightarrow$ sibling
Else
If ($\text{Sim}_{\text{FXST}}(n_{b_i} \rightarrow \text{child}, n_{f_j} \rightarrow \text{child}) > \delta$ & !Isleafnodesubtree ($n_{b_i} \rightarrow$ child, $n_{f_j} \rightarrow$ child))
IFXD($n_{b_i} \rightarrow \text{child}, n_{f_j} \rightarrow \text{child}$) //Integrate child subtree of n_{b_i}, n_{f_j}
Else //n_{f_j} added as a sibling of n_{b_i}
{InsertSub-tree($n_{b_i} \rightarrow$ common parent, n_{f_j}); deltreesubtree(n_{f_j});}
Else
If (n_{b_i} exist sibling)
IFXD($n_{b_i} \rightarrow \text{sibling}, n_{f_j}$) //Integrate sibling subtree of $n_{b_i}, n_{b_i} \rightarrow$ sibling
Else
{InsertSub-tree($n_{b_i} \rightarrow \text{parent}, n_{f_j}$); deltreesubtree($n_{f_j}$);} //$n_{f_j}$ added as a sibling of n_{b_i}
}
}
}
$FD_t = \text{Transform}\ (FT_b)$ //FXTM tree is converted to fuzzy XML document
Return FD_t

الگوریتم 2 : IFXD

6

بخش ششم

بخش ششم

ارزیابی های تجربی

- معیارهای ارزیابی
- دیتاست ها
  - دیتاست های واقعی
- نتایج تجربی

بخش ششم

معیارهای ارزیابی

$$Correctness = \frac{A}{A + B}$$

$$Completeness = \frac{A}{A + C}$$

$$F - measure = \frac{2 \times Correctness \times Completeness}{Correctness + Completeness}$$

بخش ششم

Table 1
Characteristics of the synthetic data sets.

Domain	DTD	Number of documents	Number of elements	Number of fuzzy elements	Average depth	Data set
Auction	ebay.dtd	50	156	15	3.75641	D_1
	ubid.dtd	50	342	30	3.76608	
	yahoo.dtd	50	342	30	3.76608	
University	reed.dtd	50	10546	1000	3.19979	D_2
	uwm.dtd	50	66729	5000	3.95243	
	wsu.dtd	50	74557	10000	3.15787	

جدول 1 : ویژگی های مجموعه داده های مصنوعی

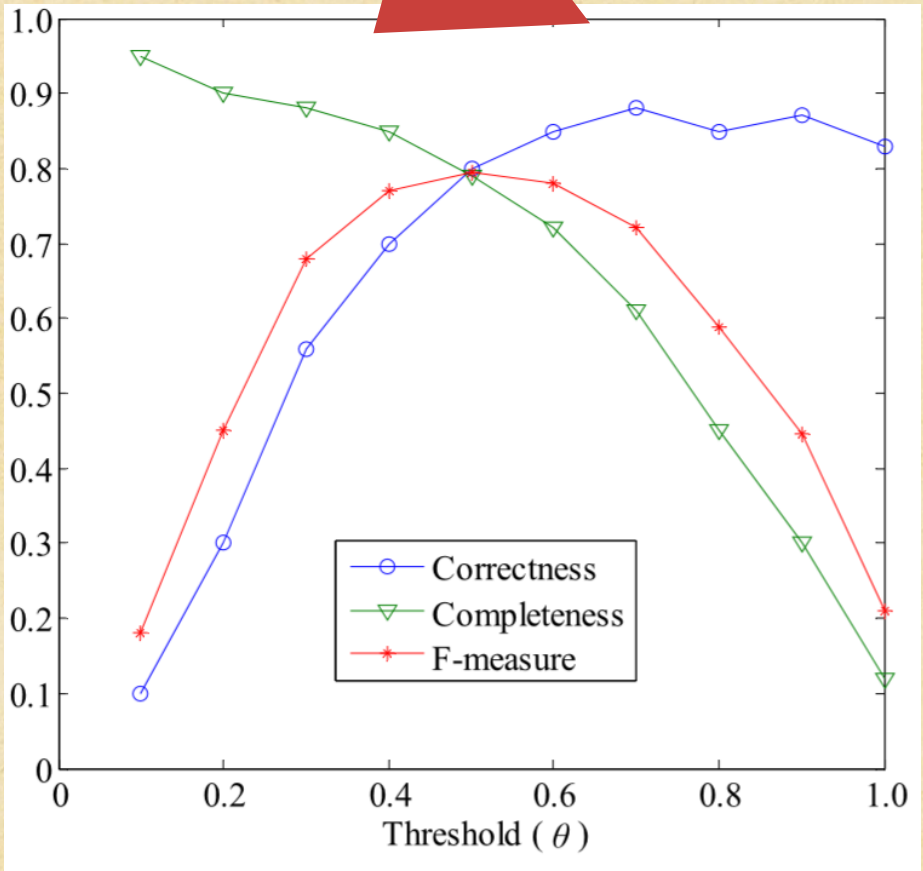
بخش ششم

Table 2
Characteristics of the real data sets.

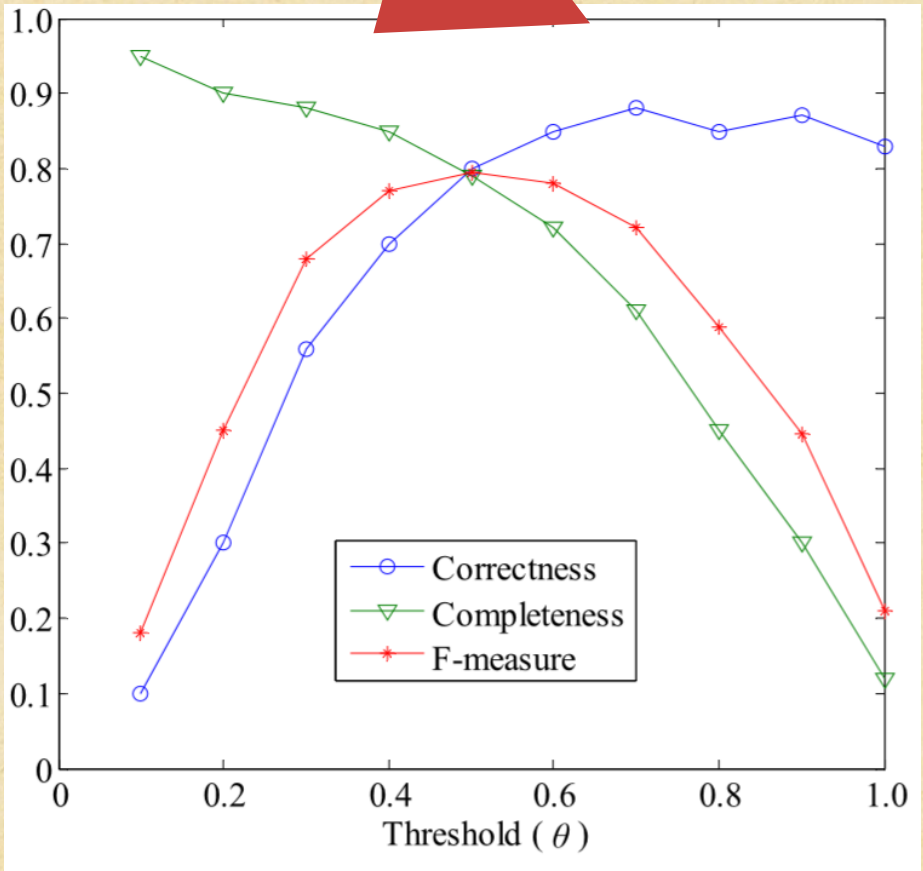
File name	Number of documents	Number of elements	Number of fuzzy elements	Average depth	Data set
NASA.xml	150	476646	5000	5.58314	D_3
DBLP.xml	150	3332130	10000	2.90228	D_4

جدول 2 : ویژگی های مجموعه داده های واقعی

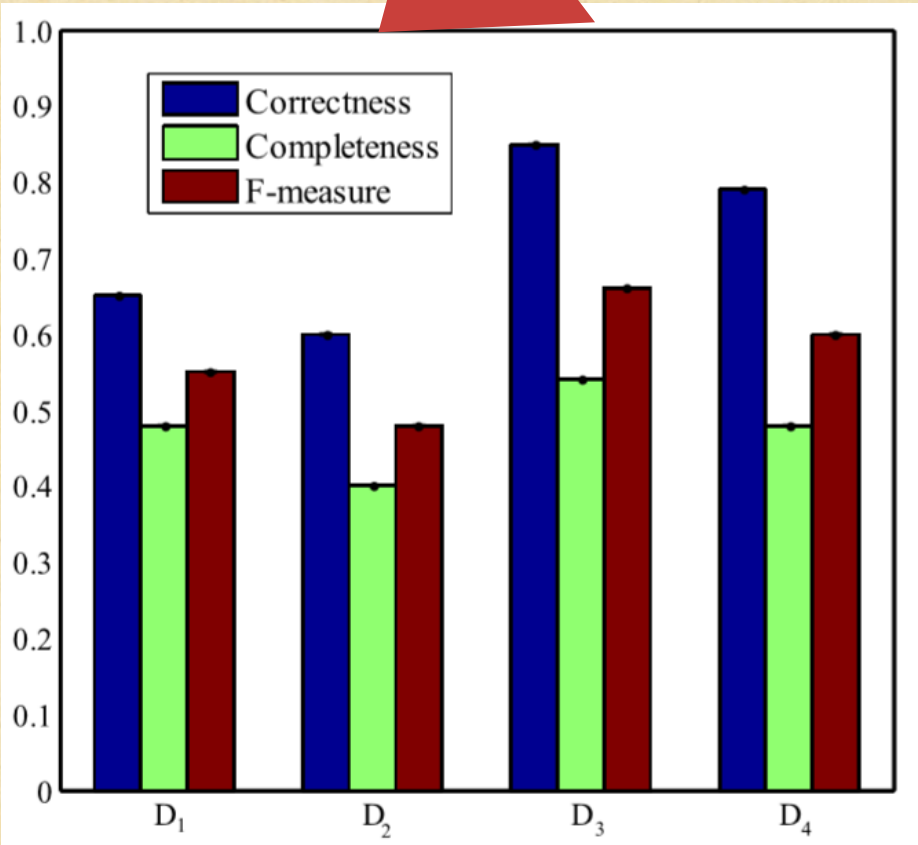
بخش ششم



شکل 6 : کیفیت داده های یکپارچه بر اساس θ

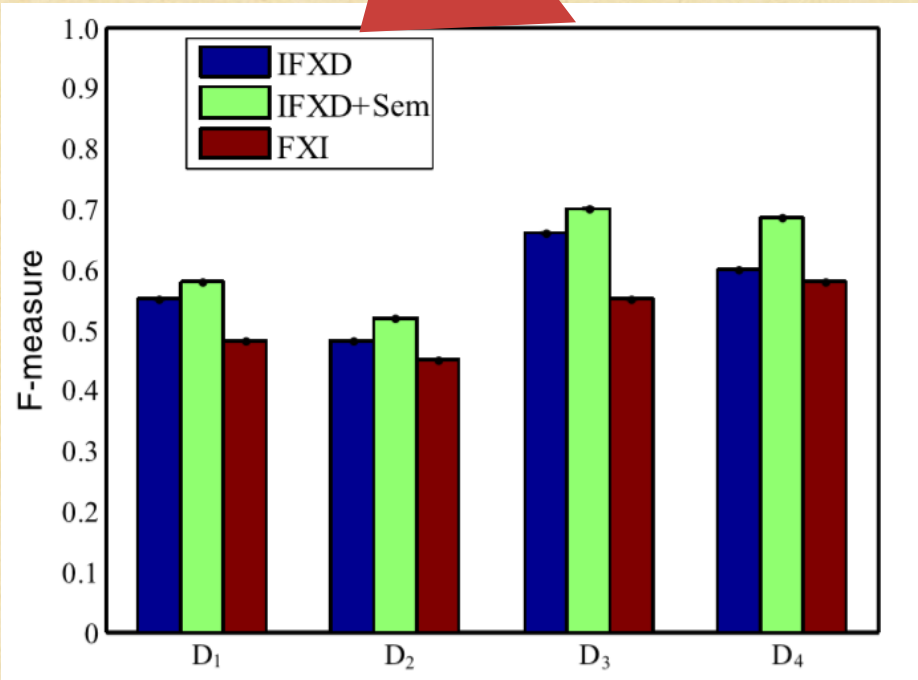


شکل 7 : کیفیت داده های یکپارچه بر اساس δ



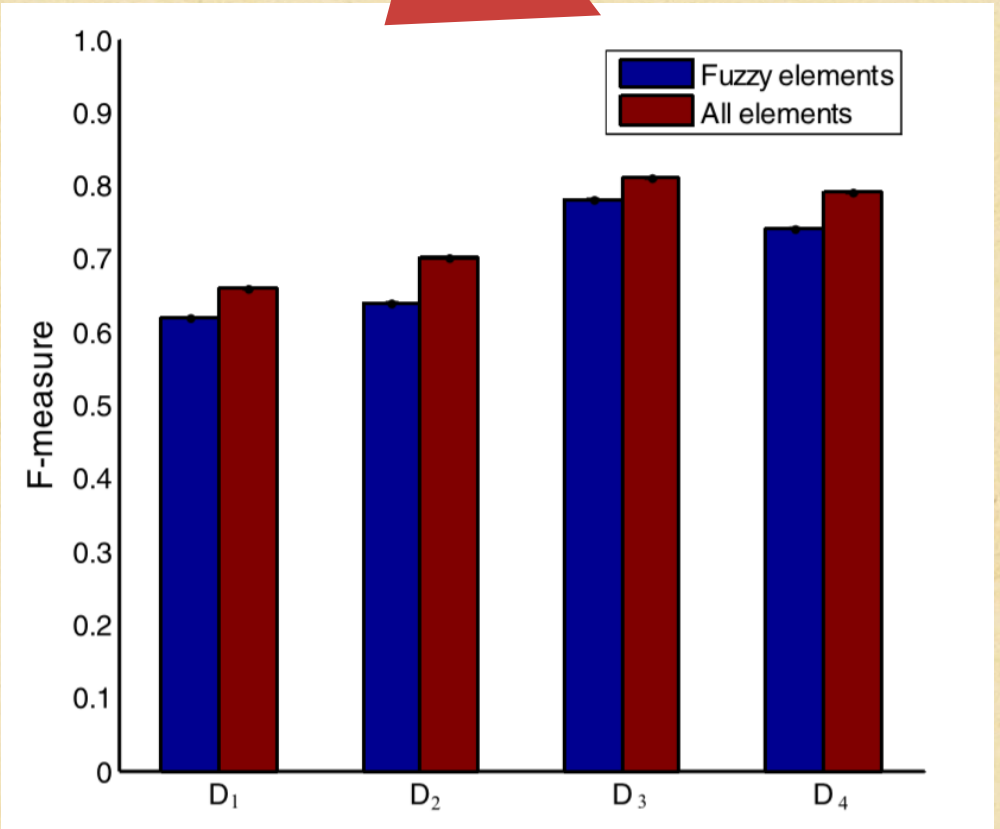
شکل 8 : مقایسه کیفیت ادغام داده ها با مجموعه های مختلف داده

بخش ششم

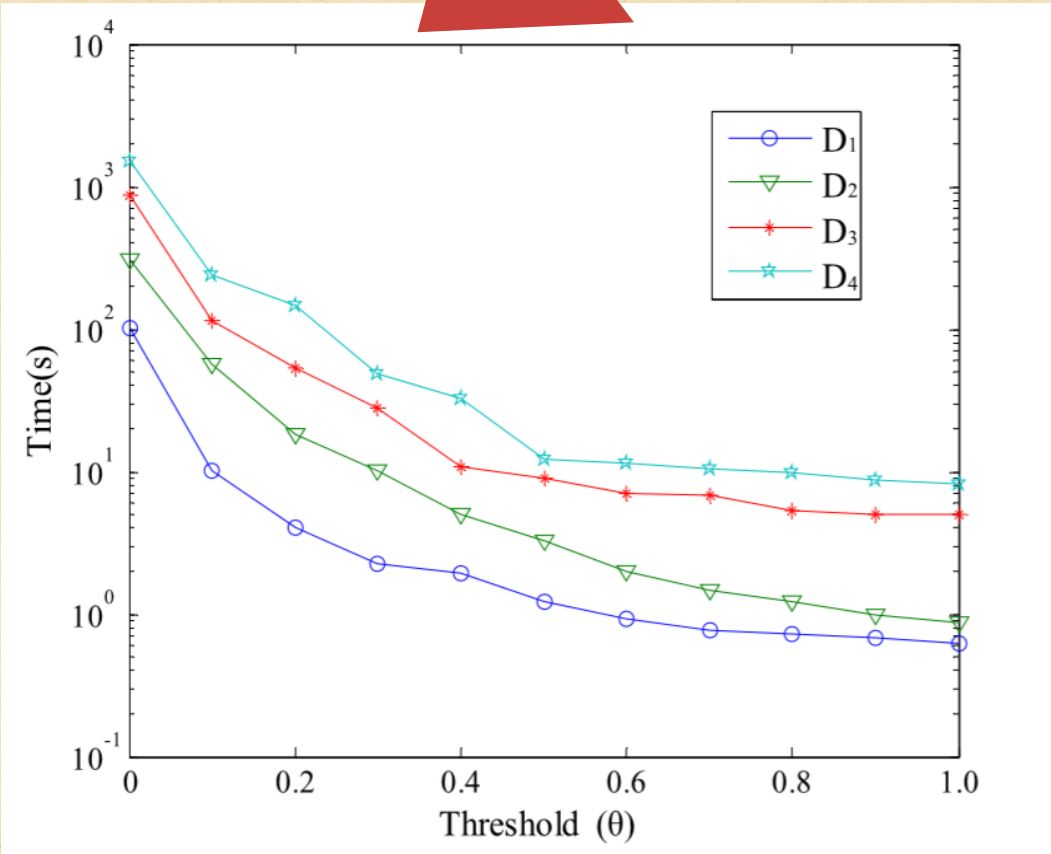


شکل 9 : مقایسه کیفیت ادغام داده ها با رویکردهای مختلف

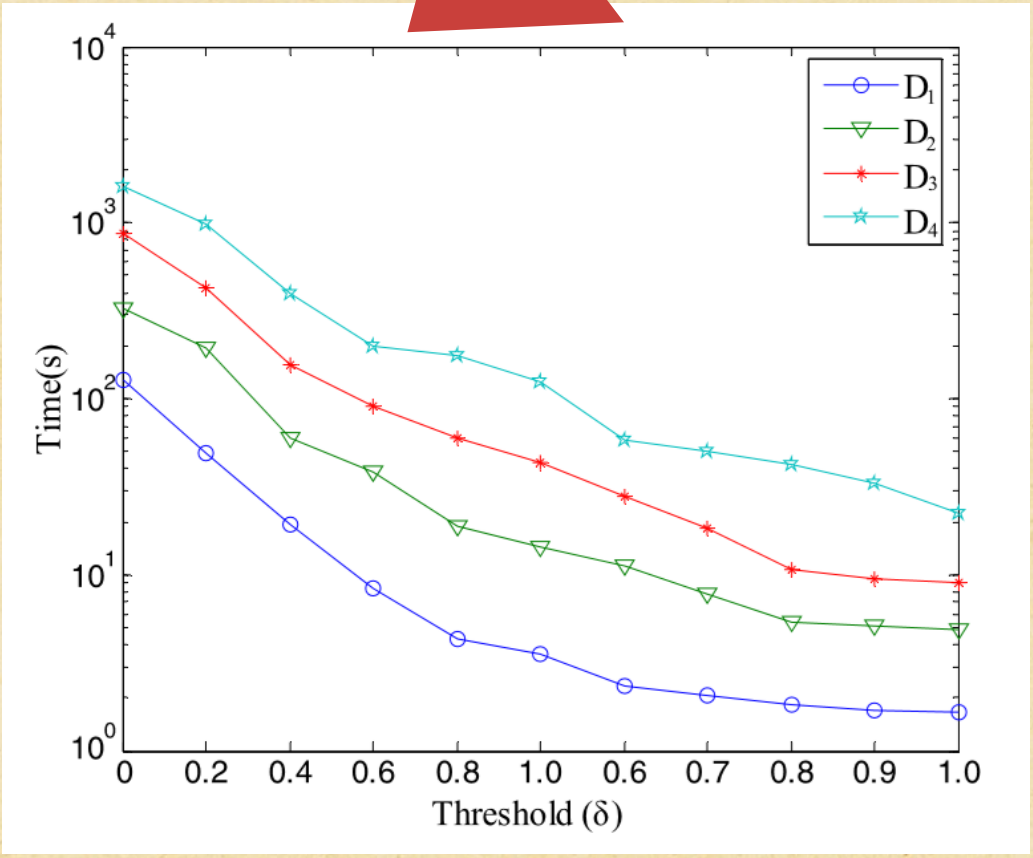
بخش ششم



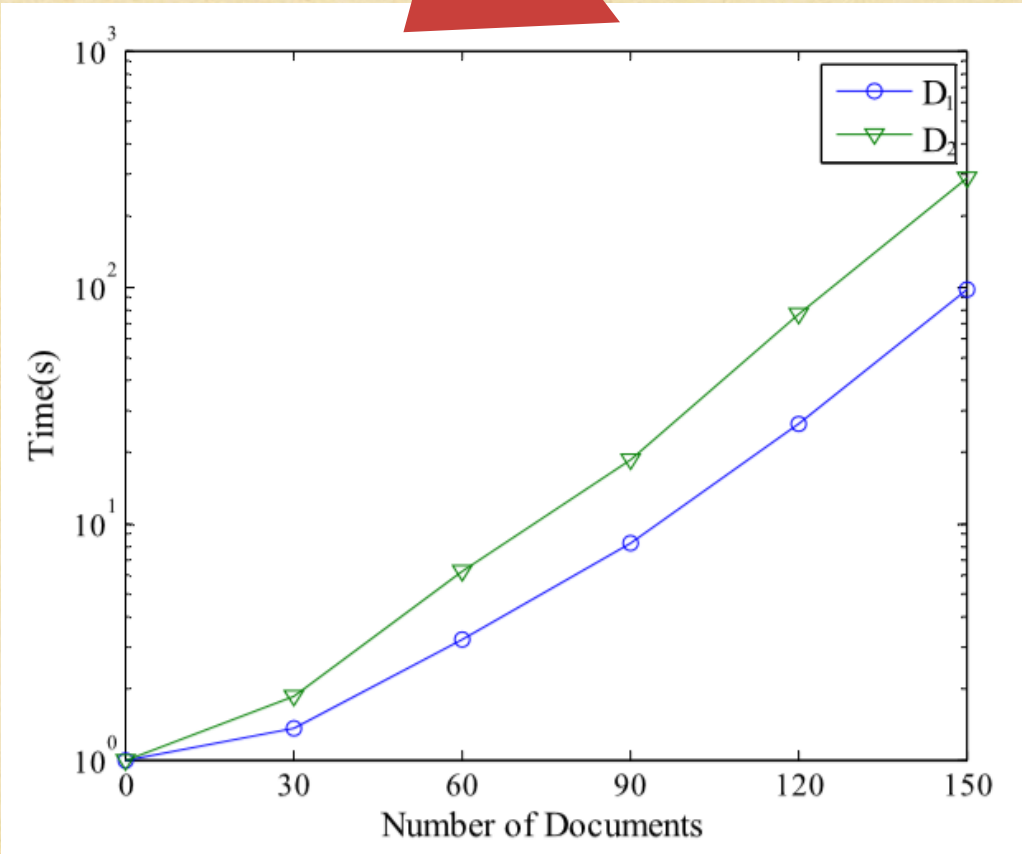
شکل 10 : مقایسه کیفیت داده های یکپارچه در سطح عناصر



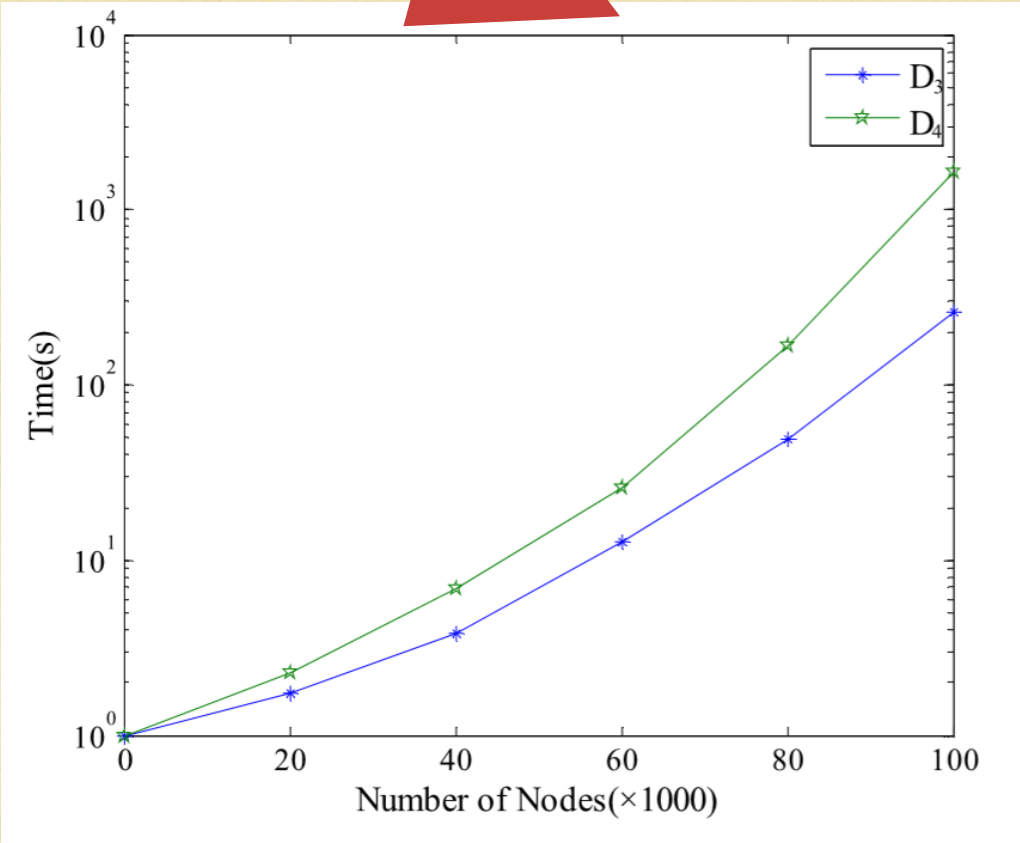
شکل 11 : زمان اجرای مجموعه داده های مختلف و آستانه θ



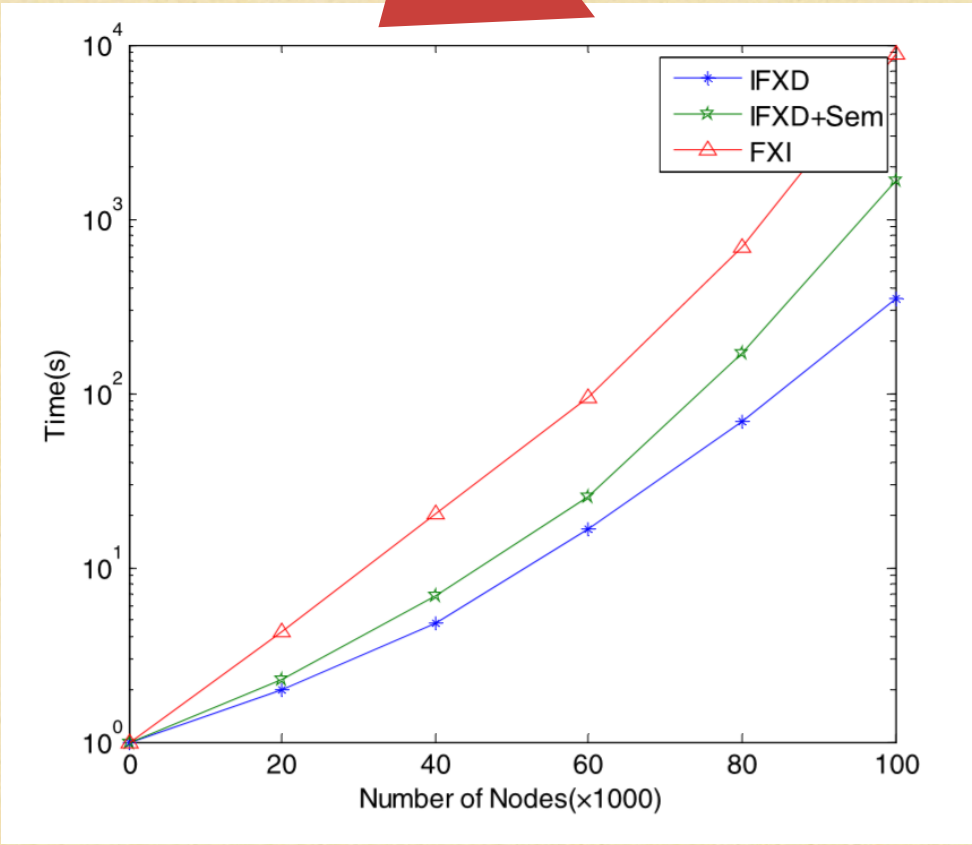
شکل 12 : زمان اجرای مجموعه داده های مختلف و آستانه δ



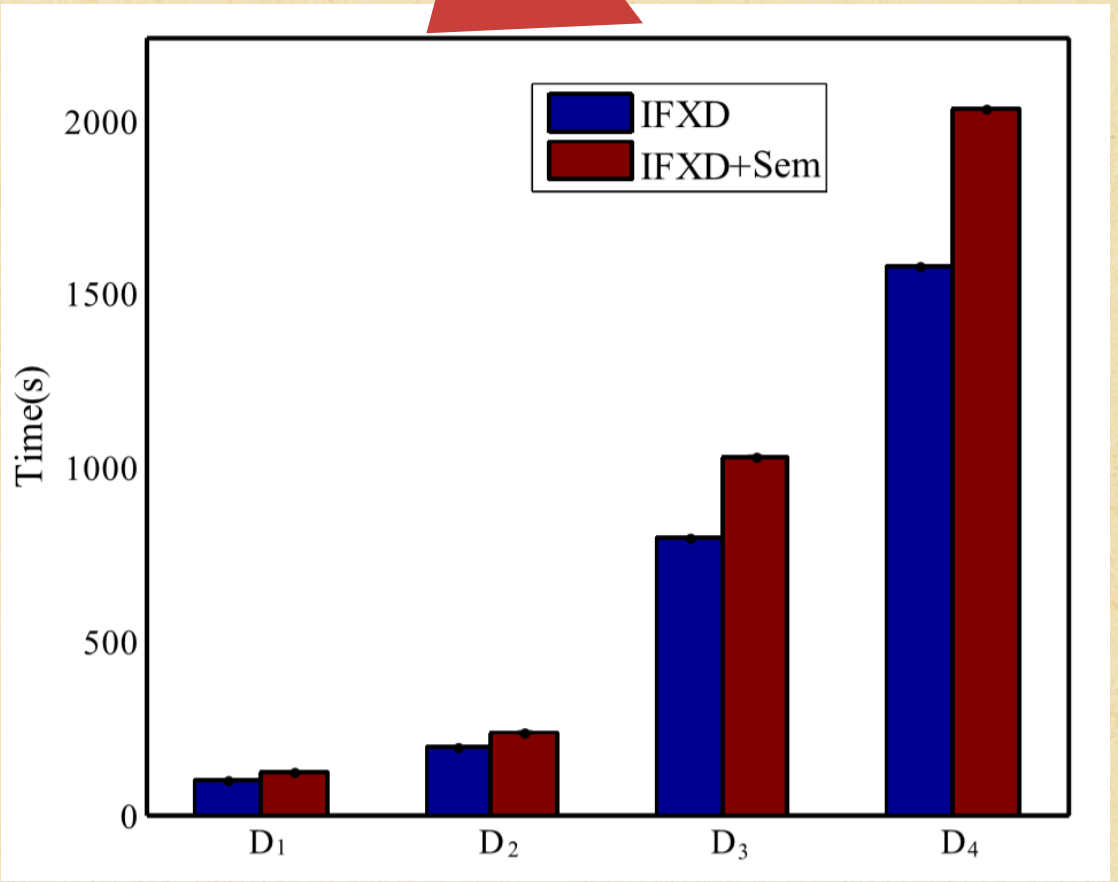
شکل 13 : مقایسه زمان ادغام برای مجموعه های مصنوعی



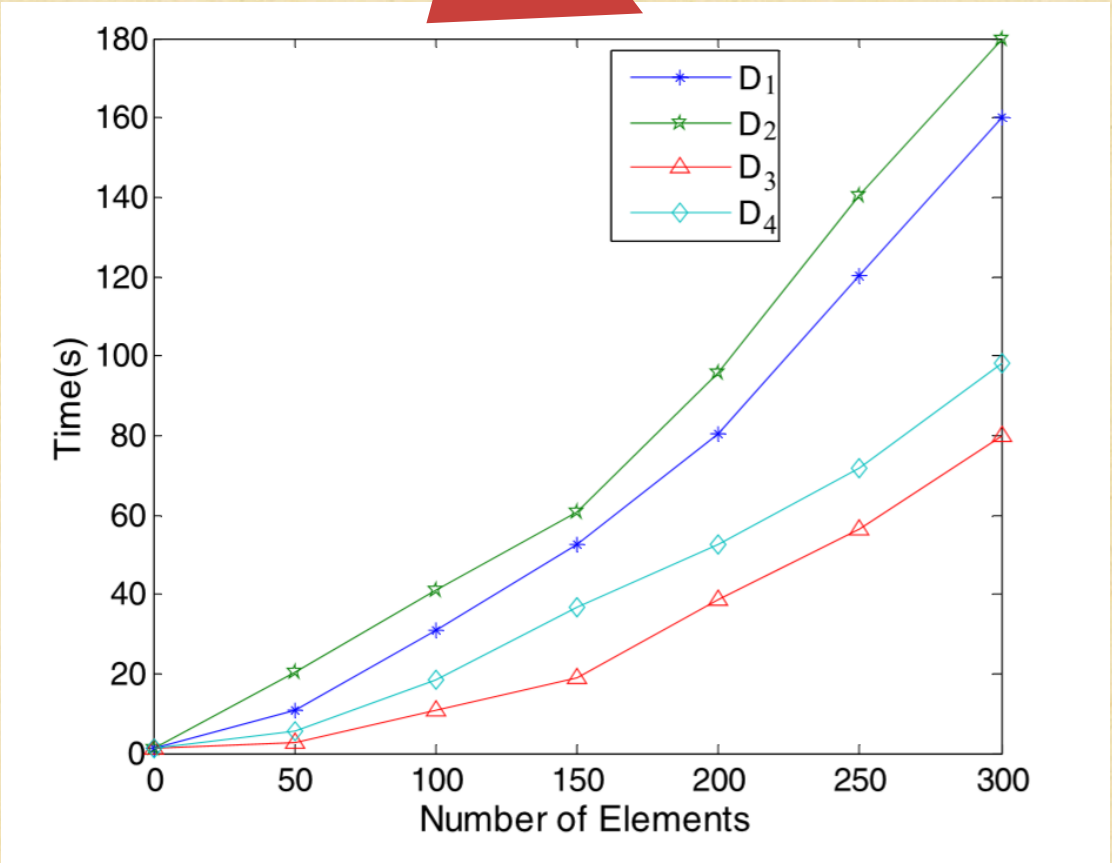
شکل 14 : مقایسه زمان ادغام برای مجموعه داده های واقعی



شکل 15 : مقایسه زمان اجرا برای استراتژی های مختلف



شکل 16 : مقایسه زمان اجرای متوسط برای دیتاست های مختلف



شکل 17 : مقایسه زمان اجرای متوسط برای تعداد مختلف عناصر

7

بخش هفتم

بخش هفتم

نتیجه گیری

برای مقابله مؤثر با ادغام داده های اسناد XMLفازی ، در این مقاله ، ما یک مدل درخت XML فازی جدید به نام مدل درخت فازی ((FXTMبرای گرفتن اطلاعات ساختاری و معنایی اسناد XMLفازی پیشنهاد کردیم. بر اساس FXTM، ما یک روش اندازه گیری شباهت شامل ساختارهای زیرسطحی دو لایه را پیشنهاد کردیم. علاوه بر این ، ما یک چارچوب برای ادغام داده های XMLفازی ناهمگن ارائه کردیم. نتایج تجربی نشان می دهد که رویکرد ما می تواند یکپارچه سازی سند XMLفازی را انجام دهد. تازگی رویکرد ما در استفاده از اندازه گیری شباهت از زیر درختان دو لایه است که برای اسناد XMLفازی اعمال می شود.

ما چندین موضوع را که باید در ادغام سند XMLفازی بررسی شود شناسایی کردیم. اولین مسئله این است که در هنگام ایجاد سند XMLفازی یکپارچه ، در نظر گرفتن محدودیت های شماتیک (به عنوان مثال ، محدودیت های کاردینالیتی و محدودیت های یکپارچگی).

بخش هفتم

نتیجه گیری

کار آینده الگوریتم شناسایی هویت را برای شرایطی که محدودیت های شماتیک در زیر لایه های دو لایه رخ می دهد، گسترش می دهد. مسئله دوم در نظر گرفتن تنظیم خودکار آستانه ها و وزن های ارائه شده توسط کاربر است به گونه ای که می توان نتایج یکپارچه سازی بهتری را برای یک دامنه کاربرد واقعی به دست آورد.

سرانجام ، در چارچوب ادغام پیشنهادی در این مقاله فرض بر این است که هر یک از منابع داده در حال ادغام تنها یک سند XMLفازی واحد هستند. این امکان وجود دارد که یک منبع داده یکپارچه از چندین سند XMLفازی مشابه شود. در این مرحله ، رویکرد پیشنهادی ادغام داده های XMLفازی ناهمگن باید برای رسیدگی به چنین مشکلی گسترش یابد.

8

بخش هشتم

بخش هشتم

منابع

- [1] L. Zamboulis, XML data integration by graph restructuring, in: Proceedings of the 2004 British National Conference on Databases, 2004, pp. 57–71.
- [2] F. Tseng, XML-based heterogeneous database integration for data warehouse creation, in: Proceedings of the 2005 Pacific-Asia Conference on Information Systems, 2005, p. 48.
- [3] A. Thomo, S. Venkatesh, Rewriting of visibly pushdown languages for XML data integration, Theor. Comput. Sci. 412 (39) (2011) 5285–5297.
- [4] N. Bikakis, et al., The XML and semantic web worlds: technologies, interoperability and integration: a survey of the state of the art, in: Semantic Hyper/Multi-media Adaptation, Springer, Berlin, 2013, pp. 319–360.
- [5] S. Abiteboul, L. Segoufin, V. Vianu, Representing and querying XML with incomplete information, ACM Trans. Database Syst. 31 (1) (2006) 208–254.
- [6] A. Nierrman, H.V. Jagadish, ProTDB: probabilistic data in XML, in: Proceedings of the 28th International Conference on Very Large Data Bases, 2002, pp. 646–657.

بخش هشتم

منابع

- [7] A.D. Keijzer, Data integration using uncertain XML, in: Soft Computing in XML Data Management, Springer, Berlin, 2010, pp. 79–103.
- [8] E. Hung, L. Getoor, V.S. Subrahmanian, PXML: a probabilistic semistructured data model and algebra, in: Proceedings of the 19th IEEE International Conference on Data Engineering, 2003, pp. 467–478.
- [9] B. Kimelfeld, Y. Kosharovski, Y. Sagiv, Query efficiency in probabilistic XML models, in: Proceedings of the 2008 ACM SIGMOD Interna-tional Conference on Management of Data, 2008, pp. 701–714.
- [10] S. Abiteboul, B. Kimelfeld, Y. Sagiv, P. Senellart, On the expressiveness of probabilistic XML models, VLDB J. 18 (5) (2009) 1041–1064.
- [11] M.V. Keulen, A.D. Keijzer, Qualitative effects of knowledge rules and user feedback in probabilistic data integration, VLDB J. 18 (5) (2009) 1191–1217.
- [12] B. Kimelfeld, Y. Kosharovsky, Y. Sagiv, Query evaluation over probabilistic XML, VLDB J. 18 (5) (2009) 1117–1140.

بخش هشتم

منابع

- [13] B. Kimelfeld, P. Senellart, Probabilistic XML: models and complexity, in: Advances in Probabilistic Databases for Uncertain Information Management, Springer, Berlin, 2013, pp. 39–66.
- [14] A. Gaurav, R. Alhajj, Incorporating fuzziness in XML and mapping fuzzy relational data into fuzzy XML, in: Proceedings of the 2006 ACM Symposium on Applied Computing, 2006, pp. 456 460.
- [15] K. Turowski, U. Weng, Representing and processing fuzzy information – an XML-based approach, Knowl.-Based Syst. 15 (1) (2002) 67–75.
- [16] B. Oliboni, G. Pozzani, Representing fuzzy information by using XML schema, in: Proceedings of the 19th International Conference on Database and Expert Systems Application, 2008, pp. 683–687.
- [17] J. Lee, Y.-Y. Fanjiang, Modeling imprecise requirements with XML, Inf. Softw. Technol. 45 (7) (2002) 445–460.
- [18] Z.M. Ma, L. Yan, Fuzzy XML data modeling with the UML and relational data models, Data Knowl. Eng. 63 (3) (2007) 972–996.

بخش هشتم

منابع

- [19] L. Yan, Z.M. Ma, F. Zhang, Fuzzy XML Data Management, Springer, Berlin, 2014.
- [20] G. Panic, M. Rackovic, S. Skrbic, Fuzzy XML with implementation, in: Proceedings of the 2012 Balkan Conference in Informatics, 2012, pp.58–62.
- [21] G. Panic,´ M. Rackovic,´ S. Škrbic,´ Fuzzy XML and prioritized fuzzy XQuery with implementation, J. Intell. Fuzzy Syst. 26 (1) (2014) 303–316.
- [22] X. Yang, M.L. Lee, T.W. Ling, Resolving structural conflicts in the integration of XML schemas: a semantic approach, in: Proceedings of the 2003 International Conference on Conceptual Modeling, Springer, Berlin, 2003, pp. 520–533.
- [23] H. Köpcke, E. Rahm, Frameworks for entity matching: a comparison, Data Knowl. Eng. 69 (2) (2010) 197–210.
- [24] X.L. Zhang, T. Yang, B.Q. Fan, Novel method for measuring structure and semantic similarity of XML documents based on extended adja-cency matrix, Phys. Proc. 24 (2012) 1452–1461.

بخش هشتم

منابع

- [25] A. Nierman, H.V. Jagadish, Evaluating structural similarity in XML documents, in: Proceedings of the 5th International Workshop on the Web and Databases, 2002, pp. 61–66.
- [26] A. Poggi, S. Abiteboul, XML data integration with identification, in: International Workshop on Database Programming Languages, 2005, pp. 106–121.
- [27] W. Liang, H. Yokota, LAX: an efficient approximate XML join based on clustered leaf nodes for XML data integration, in: British National Conference on Databases, 2005, pp. 82–97.
- [28] L. Ribeiro, T. Härder, Entity identification in XML documents, Grundl. Datenbanken (2006) 130–134.
- [29] Y. Qi, et al., XML data integration: merging, query processing and conflict resolution, in: Advanced Applications and Structures in XML Processing: Label Streams, Semantics Utilization and Data Query Technologies, IGI Global, Hershey, 2010, pp. 333–360.

بخش هشتم

منابع

- [30] S. Agreste, P.D. Meo, E. Ferrara, D. Ursino, XML matchers: approaches and challenges, Knowl.-Based Syst. 66 (2014) 190–209.
- [31] J. Tekli, R. Chbeir, A novel XML document structure comparison framework based-on sub-tree commonalities and label semantics, J. Web Semant. 11 (2012) 14–40.
- [32] J. Tekli, et al., Approximate XML structure validation based on document–grammar tree similarity, Inf. Sci. 295 (2015) 258–302.
- [33] D.D.B. Saccol, C.A. Heuser, Integration of XML data, in: Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web, Springer, Berlin, 2003, pp. 68–80.
- [34] A.M.D. Nascimento, C.S. Hara, A model for XML instance level integration, in: Proceedings of the 23rd Brazilian Symposium on Databases, Sociedade Brasileira de Computação, Porto Alegre, 2008, pp. 46–60.

بخش هشتم

منابع

- [35] A.M. Kade, C.A. Heuser, Matching XML documents in highly dynamic applications, in: Proceedings of the 8th ACM Symposium on Docu-ment Engineering, 2008, pp. 191–198.
- [36] M. van Keulen, A. de Keijzer, W. Alink, A probabilistic XML approach to data integration, in: Proceedings of the 21st International Conference on Data Engineering, 2005, pp. 459–470.
- [37] T. Pankowski, Reconciling inconsistent data in probabilistic XML data integration, in: Proceedings of the 2008 British National Conference on Databases, 2008, pp. 75–86.
- [38] A. Hamissi, B.B. Yaghlane, Belief integration approach of uncertain XML documents, in: Proceedings of IPMU'08, 2008, pp. 370–377.
- [39] M.L. Ba, et al., Integration of web sources under uncertainty and dependencies using probabilistic XML, in: Proceedings of the 2014 Interna-tional Conference on Database Systems for Advanced Applications, 2014, pp. 360–375.

بخش هشتم

منابع

- [40] J. Liu, X.X. Zhang, Data integration in fuzzy XML documents, Inf. Sci. 280 (2014) 82–97.
- [41] L. Yan, Z.M. Ma, J. Liu, Fuzzy data modeling based on XML schema, in: Proceedings of the 2009 ACM Symposium on Applied Computing, 2009, pp. 1563–1567.
- [42] G. Nicol, et al., Document object model (DOM) level 3 core specification, W3C Working Draft 13 (2001) 1–146.
- [43] R.A. Wagner, M.J. Fisher, The string-to-string correction problem, J. ACM 21 (1) (1974) 168–173.
- [44] W. Cohen, P. Ravikumar, S. Fienberg, A comparison of string metrics for matching names and records, in: KDD Workshop on Data Cleaning and Object Consolidation, 2003, pp. 73–78.
- [45] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, Sov. Phys. Dokl. 10 (8) (1966) 707–710.

بخش هشتم

منابع

- [46] G. Navarro, A guided tour to approximate string matching, ACM Comput. Surv. 33 (1) (2001) 31–88.
- [47] J.J. Jiang, D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: Proceedings of the International Conference on Research in Computational Linguistics, 1997.
- [48] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the International Joint Conference on Artificial Intelligence, 1995, pp. 448–453.
- [49] G.A. Miller, WordNet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41.
- [50] A. Marie, A. Gal, Boosting schema matchers, in: Proceedings of the OTM 2008 Confederated International Conferences, 2008, pp. 283–300.
- [51] S. Madria, K. Passi, S. Bhowmick, An XML schema integration and query mechanism system, Data Knowl. Eng. 65 (2) (2008) 266–303.

بخش هشتم

منابع

- [52] Z.M. Ma, L. Yan, Conflicts and their resolutions in fuzzy relational multidatabases, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 18 (2) (2010) 169–195.
- [53] L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets Syst. 1 (1) (1978) 3–28.
- [54] T. Dalamagas, et al., A methodology for clustering XML documents by structure, Inf. Syst. 31 (3) (2006) 187–228.
- [55] Z. Ma, L. Yan, Modeling fuzzy data with XML: a survey, Fuzzy Sets Syst. 301 (2016) 146–159.