علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۶، شماره ۱، ۱۳۹۷
صفحات ۹- ۱۶
مقاله عادی

نشریه علوم رایانش و فناوری اطلاعات

بروزرسانی و انطباق معنایی نسخ سیستم‌های پایگاه داده

محمد ظهیری دستجردی حمیدرضا احمدی‌فر

دانشکده فنی و مهندسی، دانشگاه گیلان، رشت، ایران

چکیده

در این مقاله با استناد به مقاله‌ها و رساله‌های ارائه شده تا سال ۲۰۱۶، پرکاربردترین روش‌ها و متدهایی که جهت انطباق اجزای نسخ اولیه و هدف یک پایگاه داده مورد استفاده قرار می‌گیرند، همراه با معایب و مزایای آنها مورد بررسی قرار گرفته است. در ادامه به صورت ویژه اشکالات انطباق براساس تکنیک سنجش شباهت رشته‌ای و آن هم با استفاده از تغییرات در نسخ میانی بررسی گردیده و در نهایت راه حلی ارایه شده، که هم بتوان از اطلاعات مربوط به تغییرات اجزاء در نسخ میانی استفاده نمود و هم سنجش شباهت معنایی را در کنار سنجش شباهت رشته‌ای مورد استفاده قرار داد تا احتمال خطا به میزان چشم‌گیری کاهش پیدا کند. برای این کار از دو جدول شباهت مجزا و مستقل که براساس شباهت رشته‌ای و معنایی تولید شده‌اند و یک جدول شباهت نهایی که عناصرش ترکیبی از دو جدول قبلی است استفاده شده است. در پایان در محیطی یکسان این روش با روش‌هایی که تطبیق‌گرهای معروف از آن استفاده می‌کنند مقایسه گردیده است. نتایج حاصل از آزمایشات نشان می‌دهد که روش پیشنهادی در اکثر موارد از دقت و صحت بالاتری در عملیات انطباق برخوردار است.

کلمات کلیدی: انطباق، تطبیق‌گر، پایگاه داده، سنجش شباهت معنایی.

۱- مقدمه

امروزه شاهد آن هستیم که برنامه‌نویسان و توسعه‌دهندگان در حوزه‌های مختلف، بنا بر نیازهای جدید کاربران خود اقدام به بروزرسانی[۱] نرم‌افزارها می‌کنند. در این راستا دو بحث مهم قابل طرح و بررسی است. یکی بروزرسانی نرم‌افزارها به‌صورت پویا و برخط [۱] و دیگری انطباق[۲] و بروزرسانی نسخه‌های مختلف پایگاه داده که در برنامه‌ها استفاده می‌گردند. انطباق و بروزرسانی دو فرآیند کاملا مستقل می‌باشند. انطباق مقدمه بروزرسانی است. در فرآیند انطباق ممکن است با چالش‌ها و مشکلات زیادی مواجه شویم تا جایی که اجرای موفقیت‌آمیز فرآیند را غیرممکن کنند. یکی از این مشکلات ناهمگونی معنایی بالایی است که در بعضی طرح‌های پایگاه داده[۳] وجود دارد. به‌عنوان مثال توسعه‌دهندگان سیستم‌های پایگاه داده، عناصر پایگاه را براساس یک نظم ساختاری که هیچ رابطه معنایی با یکدیگر نداشته باشند، نام‌گذاری کنند. نام‌هایی همچون ver1t1, ver1t2, … ver2t1, … ver2t2, که به‌عنوان اسامی جداول در طرح‌های مختلف یک پایگاه داده در نظر گرفته شده است، می‌تواند توسط الگوریتم‌های موجود که فقط شباهت‌های رشته‌ای را تشخیص می‌دهند به‌عنوان اسامی منطبق شناسایی شود، در صورتی‌که ممکن است هیچ قرابت معنایی بین آنها وجود نداشته باشد. وقوع چنین شرایطی منجر به تولید نتایجی غیرواقعی در انواع طرح‌های انطباق می‌شود.

در طی اجرای فرآیند طرح‌های انطباق، شباهت‌های رشته‌ای[۴] و معنایی[۵] که به‌صورت اتوماتیک تشخیص داده می‌شوند ممکن است نیازمند دخل و تصرف و ویرایش توسط کاربران باشند (حذف ارتباط‌های اشتباه، اضافه کردن ارتباط‌های از قلم افتاده و غیره)، این عملیات تا جایی انجام می‌شود که نگاشت صحیح حاصل شود [۲]. اما یکی از مواردی که این ابزارها را می‌تواند از همدیگر متمایز نماید میزان دخل و تصرف کاربران است که سعی شده است در روش پیشنهادی به حداقل ممکن برسد.

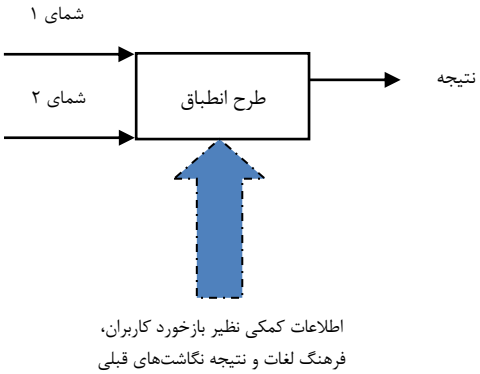
نکته حائز اهمیت آن است که هیچ فرآیند انطباق تاکنون نتوانسته است به‌صورت صد در صد و دقیق نیازهای کاربران را برآورده کند. بلکه تطبیق‌گرها[۶] سعی نموده‌اند با استفاده از روش‌های ترکیبی، انطباق را دقیق‌تر و صحیح‌تر انجام دهند.

در مبحث انطباق، الگوریتم‌ها و نظریه‌های زیادی ارائه شده است. که مهم‌ترین

آن‌ها در مقاله Euzenat and Shvaiko در سال ۲۰۱۳ مورد بررسی قرار گرفته است [۳].

۲- مفهوم انطباق

انطباق در حقیقت شناخت هر نوع ارتباط معنایی و یا کشف شباهت ساختاری بین عناصر دو طرح یک پایگاه داده است. به نحوی که کشف و استنتاج این روابط بتواند به ما در عملیات انتقال، بروزرسانی و هر نوع ارتباط داده‌ای دیگر کمک کند [۴]. شکل ۱ نمای ساده‌ای از فرآیند انطباق را نشان می‌دهد:

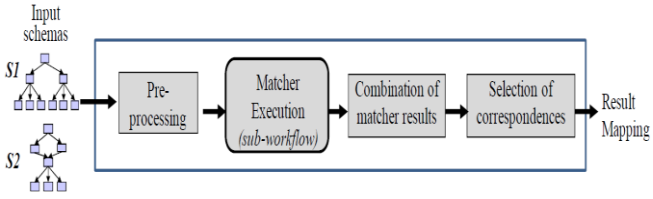


شکل ۱- نمای ساده‌ای از چگونگی فرآیند انطباق [۴]

در شکل ۱ دو طرح قدیمی و جدید از یک پایگاه داده به همراه اطلاعات کمکی به یک طرح انطباق داده می‌شود تا پس از عملیات پردازشی، نتیجه که در حقیقت لیستی از ارتباطات معنایی و شباهت‌های ساختاری است تولید گردد. اطلاعات کمکی دربرگیرنده بازخورد کاربر (توسعه‌دهندگان)، واژه‌نامه و نتیجه انطباق‌های قبلی است.

۲-۱- روند کاری عملیات انطباق

شکل ۲ روند کاری عملیات انطباق را که به‌صورت عمومی در ابزارهای انطباق مورد استفاده قرار گرفته است، تشریح می‌کند [۲].



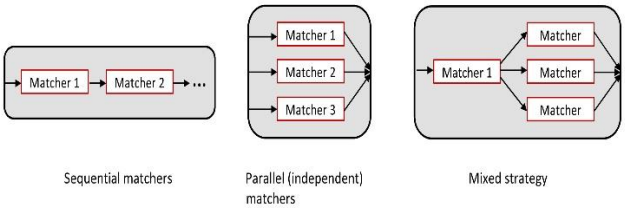
شکل ۲- روند عمومی عملیات انطباق در روش‌های کنونی [۲]

در طرح انطباق شکل ۲ چهار مرحله به‌صورت سریال تعریف شده است:

۱. اولین مرحله پیش پردازش[۷] است. در این قسمت عملیاتی نظیر واکشی اطلاعات عناصر پایگاه‌ها، ایندکس‌گذاری نام‌های صفات و خصیصه‌ها جهت تسریع بخشیدن به عملیات تعیین اسامی متشابه[۸] مورد استفاده قرار می‌گیرد [۵].

۲. دومین مرحله شامل مجموعه‌ای از جریان‌های کاری[۹] است. این مرحله اصلی‌ترین قسمت فرآیند انطباق را دربر می‌گیرد.

در این مرحله انواع تطبیق‌گرها براساس الگوریتم‌های مختلف اجرا گردیده و نتایج متفاوتی را تولید می‌کنند. این نتایج در مراحل بعدی پایه و اساس انطباق را تشکیل می‌دهند. تطبیق‌گرها می‌توانند نسبت به یکدیگر در حالات مختلفی اجرا گردند. مجموعه‌ای از این حالات در شکل ۳ نشان داده شده است [۲].



شکل ۳- حالت‌های مختلف قرارگیری تطبیق‌گرها در Sub Workflow [۲]

همانطور که در شکل ۳ نشان داده شده است، تطبیق‌گرها ممکن است به‌صورت سریال و به نوبت، یا به‌طور مستقل و موازی و یا در حالت ترکیبی اجرا شوند [۶]. در حالت سریال، تطبیق‌گرها به‌صورت مستقل از هم اجرا نمی‌شوند بلکه نتایج هر تطبیق‌گر به‌عنوان ورودی تطبیق‌گر بعدی مورد استفاده قرار می‌گیرد. در حالت موازی، تطبیق‌گرها به‌صورت مستقل از هم اجرا می‌گردند. در این حالت ما به تعداد تطبیق‌گرها نتایج متفاوت خواهیم داشت. از مزیت مهم این حالت می‌توان به اجرای هم‌زمان و موازی تطبیق‌گرها روی چند پردازنده اشاره نمود که طبیعتاً زمان اجرای طرح انطباق را به میزان محسوسی کاهش می‌دهد. در حالت ترکیبی می‌توان تطبیق‌گرها را هم به‌صورت موازی و هم به‌صورت سریال اجرا نمود و پس از آن نتایج حاصله را با هم ترکیب کرد [۲].

۳. سومین مرحله ترکیب نتایج حاصل از اجرای تطبیق‌گرها در مرحله دو و تولید نتایج ترکیبی است.

۴. چهارمین مرحله، پالایش و انتخاب نتیجه صحیح‌تر و دقیق‌تر از بین انطباق‌های استخراج شده است.

۲-۲- سطوح انطباق

از آنجایی‌که تغییرات نسخه‌های مختلف پایگاه داده می‌تواند در سطوح مختلفی صورت پذیرد، فرآیند انطباق نیز سطوح مختلفی دارد. این سطوح می‌تواند شامل سطح شمای پایگاه داده و سطح معنایی باشد. سطوح انطباق می‌تواند در هر سطح تغییرات اعمال شود [۴].

۲-۲-۱- انطباق درسطح شمای پایگاه داده

سطح شمای پایگاه داده یکی از سطوحی است که بیشترین تغییرات در آن رخ می‌دهد. تغییرات اعمال شده بر هر عنصر شامل حذف، اضافه و ویرایش می‌باشد. هر چه بازه این تغییرات گسترده‌تر باشد فرآیند انطباق نیز چالش‌های بیشتری را پیش‌رو خواهد داشت. لیستی از این تغییرات را می‌توان در جدول ۱ مشاهده نمود [۷].

سطح شمای پایگاه داده خود به دو بخش تقسیم می‌گردد:

- سطح عنصر[۱۰]

در این سطح فقط شباهت‌ها براساس مقادیر نام، توصیف، نوع داده‌ها، رابطه جداول از نوع is-a و محدودیت‌ها و ساختار، مورد سنجش واقع می‌شود.

- سطح ساختاری[۱۱]

در این سطح ترکیبی از عناصر موجود در پایگاه منبع در قالب یک ساختار با ترکیبی از عناصر در پایگاه داده هدف انطباق داده می‌شود.

جدول ۱- انواع تغییرات ممکن بین دو طرح پایگاه داده همراه با دستورات SQL [۲]

Ref.	Atomic Change	Category	DDL (MySql implementation)
A1	Add Table	Transformation	CREATE TABLE *t_name*
A2	Add Column	Transformation	ALTER TABLE *t_name* ADD *c_name*
A3	Add Veiw	Transformation	CREATE VIEW *v_name* AS..
A4	Drop Table	Structure Refactoring	DROP TABLE *t_name*
A5	Rename Table	Structure Refactoring	ALTER TABLE *o_t_name* RENAME *n_t_name*
A6	Drop Column	Structure Refactoring	ALTER TABLE *t_name* DROP COLUMN *c_name*
A7	Rename Column	Structure Refactoring	ALTER TABLE *t_name* CHANGE COLUMN *o_c_name n_c_name*
A8	Change Column Datatype	Structure Refactoring	ALTER TABLE *t_name* MODIFY COLUMN *c_name c_def*
A9	Drop View	Structure Refactoring	DROP VIEW *v_name*
A10	Add Key	Structure Refactoring	ALTER TABLE *t_name* ADD KEY *k_name*
A11	Drop Key	Structure Refactoring	ALTER TABLE *t_name* DROP KEY *k_name*
A12	Add Foreign key	Referential Integrity Refactoring	ALTER TABLE *t_name* ADD FOREIGN KEY *fk_name*..
A13	Drop Foreign key	Referential Integrity Refactoring	ALTER TABLE *t_name* DROP FOREIGN KEY *fk_name*
A14	Add Trigger	Referential Integrity Refactoring	CREATE TRIGGER *trig_name*.. ON TABLE *t_name*..
A15	Drop Trigger	Referential Integrity Refactoring	DROP TRIGGER *trig_name*
A16	Add Index	Architectural Refactoring	ALTER TABLE *t_name* ADD INDEX *idx_name*
A17	Drop Index	Architectural Refactoring	ALTER TABLE *t_name* DROP INDEX *idx_name*
A18	Add Column Default Value	Data Quality Refactoring	ALTER TABLE *t_name* MODIFY COLUMN *c_name* SET DEFAULT *value*
A19	Drop Column Default Value	Data Quality Refactoring	ALTER TABLE *t_name* MODIFY COLUMN *c_name* DROP DEFAULT
A20	Change Column Default Value	Data Quality Refactoring	ALTER TABLE *t_name* MODIFY COLUMN *c_name* SET DEFAULT *value*
A21	Make Column Not NULL	Data Quality Refactoring	ALTER TABLE *t_name* MODIFY COLUMN *c_name* NOT NULL
A22	Drop Column Not NULL	Data Quality Refactoring	ALTER TABLE *t_name* MODIFY COLUMN *c_name* NULL
A23	Add Stored Procedure	Method Refactoring	CREATE PROCEDURE *pro_name*..
A24	Drop Stored Procedure	Method Refactoring	DROP PROCEDURE *pro_name*

۲-۲-۲- انطباق معنایی[۱۲]

انطباق در این سطح براساس محتوی و معنای عناصر طرح پایگاه داده مبدا و مقصد انجام می‌شود. انطباق در این سطح وقتی مفید است که اولاً میزان اطلاعات موجود در پایگاه داده محدود باشد و ثانیاً داده‌ها نیمه ساخت یافته باشند. در این سطح می‌توان تفسیرهای نادرستی که حاصل از انطباق سطح شمای پایگاه داده است را اصلاح کرد.

در این تحقیق انطباق معنایی تحت عنوان سنجش شباهت معنایی معرفی و مورد استفاده قرار گرفته است. قابل ذکر است یکی از بزرگ‌ترین مشکلات این روش جمع‌آوری داده‌های یک پایگاه دانش به‌عنوان مبنای سنجش شباهت معنایی است [۲].

۳- تطبیق‌گرها

تاکنون تطبیق‌گرهای زیادی جهت اجرای فرآیند انطباق، ویژه محیط‌های عملیاتی مختلف طراحی و اجرا گردیده‌اند. بعضی از تطبیق‌گرها نیز استفاده‌های عمومی دارند که طبیعتاً با چالش‌های بیشتری مواجه هستند. در این بخش چند تطبیق‌گر که در مسابقات امتیازدهی OAEI[۱۳] شرکت نموده و عملکرد طرح پیشنهادی با آنها مقایسه شده است معرفی می‌گردند. OAEI یک انجمن بین‌المللی جهت ارزیابی تطبیق‌گرها و طرح‌های انطباقی است و سالانه مسابقاتی را در همین جهت برگزار می‌نماید.

در پایان این بخش یک طرح انطباق که در روش پیشنهادی از آن استفاده شده معرفی و مورد ارزیابی قرار می‌گیرد:

۳-۱- تطبیق‌گر Coma++

Coma++ یک نمونه از تطبیق‌گرها است که از معماری ترکیبی استفاده می‌کند و با یک رابط کاربر گرافیکی از محبوب‌ترین نمونه‌ها در نوع خود به شمار می‌آید. این ابزار محصول گروه پایگاه داده دانشکده علوم کامپیوتر دانشگاه Leipzig آلمان است. در این ابزار، از انطباق‌دهی در سطح شمای پایگاه داده نیز پشتیبانی می‌شود. در طراحی Coma++ از نظریه‌های خوشه‌بندی و بازیافت استفاده گردیده است.

Coma++ یک ابزاری است که به‌صورت متن‌باز توسعه یافته و در دسترس عموم قرار دارد. از جمله مهم‌ترین کاربردهای این ابزار انطباق طرح‌های XML، شبکه و مدل‌های مهندسی نرم‌افزار (UML) است [۸ و ۹].

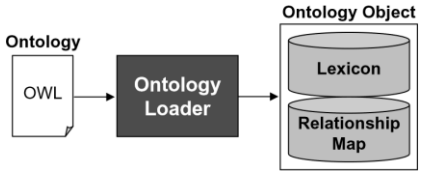
۳-۲- تطبیق‌گر Asmov

Asmov در مسابقات OAEI ۲۰۱۱ جز بهترین تطبیق‌گرها از لحاظ کارآیی بوده است. مهمترین امتیاز این ابزار پالایش نتایج در مرحله چهارم از اجرای فرآیند انطباق است. در این مرحله تناقض‌های موجود در نتایج حاصل از پردازش‌ها اصلاح و یا حذف می‌گردند. یکی از تناقض‌های مهمی که در الگوریتم‌های پالایش مورد بررسی قرار می‌گیرد، ارتباطات مورب است [۶].

۳-۳- تطبیق‌گر AML

در سال ۲۰۱۴ تطبیق‌گر AgreementMakerLight یا در اختصار AML به عنوان تطبیق‌گر ترکیبی با شرکت در مسابقات OAEI توانست عنوان نخست را کسب کند [۱۰].

در این تطبیق‌گر از یک واسط بارگزاری‌کننده[۱۴] استفاده می‌شود. این واسط طرح‌های پایگاه داده را از یک فایل با فرمت OWL دریافت نموده و سپس به دو نوع داده Lexicon و Relationship Map در قالب یک شیئ تبدیل می‌کند. این شیئ می‌تواند توسط موتور تطبیق‌گر AML مورد استفاده قرار گیرد. شکل ۴ این روند را نشان می‌دهد:



شکل ۴- ماژول بارگزاری کننده طرح از فایل OWL

۳-۴- استفاده از تاریخچه تغییرات نسخه‌ها برای تطبیق و بروزرسانی سیستم‌های پایگاه داده

یکی از تکنیک‌های انطباق، تکنیک مبتنی بر طبقه‌بندی است [۴]. این تکنیک برگرفته از الگوریتم‌های گراف است. ایده مطرح شده در این تکنیک چنین می‌گوید:

با استفاده از لینک‌های ارتباطی از نوع is-a بین گره‌های یک گراف، ممکن است بتوان براساس روابط وراثت یا روابط بین مجموعه‌ها و زیرمجموعه‌ها، رابطه‌های جدیدی را کشف و استخراج کرد [۴].

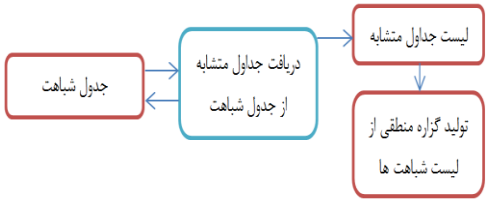
روشی که تحت عنوان "استفاده از تاریخچه تغییرات نسخه‌ها برای تطبیق و بروزرسانی سیستم‌های پایگاه داده" [۱۳] ارائه گردیده سعی دارد جهت انطباق دو نسخه از یک پایگاه داده، با بسط تکنیک طبقه‌بندی، از اطلاعات نسخه‌های میانی نیز استفاده کند و با بهره‌گیری از این اطلاعات مشابهت‌ها و روابط جدیدی را استنتاج نموده و نهایتاً در انطباق نسخه اولیه و هدف از آن‌ها بهره‌برداری کند.

در این روش اطلاعاتی که از نسخ میانی استخراج می‌شود صرفاً براساس

سنجش درصد شباهت‌های رشته‌ای بین هر دو جزء از عناصر کلیه طرح‌های شرکت‌کننده در انطباق است. این شباهت‌ها براساس درصد در جدولی تحت عنوان جدول شباهت ثبت می‌شود (مراحل ساخت جدول شباهت در بخش روش پیشنهادی تشریح شده است) که نهایتاً منجر به تولید گزاره‌های منطقی می‌شود.

۳-۴-۱- تولید گزاره‌های منطقی

بعد از ایجاد جدول شباهت گزاره‌های منطقی تحت عنوان یک حقیقت[۱۵] تولید می‌گردند. روند کار در شکل زیر نشان داده شده است.



شکل ۵- مراحل انطباق و تولید گزاره‌های منطقی [۱۳]

پس از آن توسط یکی از ابزارهای پرس و جوی منطقی مانند پرولوگ تمامی روابط به‌صورت سؤالی پرسیده می‌شود و کامپایلر با استفاده از پایگاه دانش خود و با استفاده از رابطه تعمیم به سؤالات پاسخ می‌دهد و عملیات انطباق انجام می‌گیرد.

به‌عنوان مثال مجموعه‌ای از حقایق و رول‌ها[۱۶] می‌تواند به‌صورت زیر باشد:

Equal (table 1, table 3).

Match (X, Z): -equal (X, Z).

Match (X, Z): -equal (X, Y), match (Y, Z).

نقطه قوت این روش بسط تکنیک طبقه‌بندی[۱۷] [۴] در انطباق و استفاده از اطلاعات نسخ میانی بین نسخه اولیه و هدف است. این روش سه نقطه ضعف اساسی نیز دارد:

- یکی آن که تمرکزش بر روی سنجش شباهت رشته‌ای در تولید جدول شباهت‌ها است. این موضوع علاوه بر آنکه در تناقض با روابط از نوع is-a در تکنیک طبقه‌بندی انطباق است، می‌تواند در بسیاری از محیط‌های عملیاتی که بروزرسانی نسخ پایگاه داده خود را با تمرکز بر روی تفاوت‌های معنایی استوار می‌کنند منجر به یک انطباق ضعیف گردد.
- دوم آنکه برای استنتاج از گزاره‌های منطقی از ابزارهای پرس‌وجوی منطقی[۱۸] استفاده می‌کند. این ابزارها در نهایت انطباق عنصری از پایگاه مبدأ به عنصری در پایگاه مقصد را اعلام می‌کنند و علاوه بر آنکه هیچ پالایشی بر روی استنتاج‌ها انجام نمی‌دهند تاریخچه‌ای از شرکت نسخه‌های واسط در استنتاج‌های جدید را نیز گزارش نمی‌کند.

به‌عبارت دیگر در این روش انطباق‌های موازی و چندگانه تفکیک و گزارش نخواهند شد و ممکن است یک عنصر از پایگاه مبدأ هم‌زمان به چندین عنصر در پایگاه هدف و یا بالعکس منطبق شود.

- سوم آنکه به‌دلیل استفاده از محیط‌های مجزا برای عملیات انطباق، زمان اجرای الگوریتم‌ها افزایش پیدا می‌کند. در این میان زمان قابل توجهی صرف واکشی اطلاعات از مخازن داده مشترک با محیط‌های بیرونی از جمله پرولوگ صرف می‌شود. زمانی نیز صرف سوئیچ بین ابزارهای مختلف خواهد شد.

روش پیشنهادی از روش استفاده از تاریخچه تغییرات نسخه‌ها به‌عنوان جزئی از تکنیک‌های ترکیبی استفاده کرده و نقاط ضعف آن را برطرف می‌کند.

۴- طرح مسئله

همچنان که در شکل ۱ آمده است اطلاعات کمکی می‌تواند در دقت انطباق نقش زیادی داشته باشد. پس برای داشتن انطباقی دقیق‌تر و صحیح‌تر، باید از طرح خود اطلاعات بیشتری جمع‌آوری نمود. تغییرات پایگاه داده در روند توسعه نرم‌افزار کمتر به‌صورت ناگهانی انجام می‌شود [۷] و می‌توان این تغییرات را دنبال نمود و اطلاعات کمکی مفیدی را جمع‌آوری کرد. باید توجه داشت که اگر طرح‌های پایگاه داده گوناگون و نامتقارن باشند، ارتباط معنایی و شباهت ساختاری نیز وجود نخواهد داشت و انطباق نمی‌تواند مفید باشد [۲].

روش پیشنهادی در این تحقیق برای انطباق به‌صورت ترکیبی بر سه تکنیک استوار است:

۱. استفاده از تغییرات نسخ میانی در استخراج داده‌های پایگاه دانش[۱۳].

۲. استفاده از تکنیک سنجش تشابه رشته‌ای توسط الگوریتم Jaro [۱۳]. این الگوریتم یکی از معروف‌ترین الگوریتم‌های سنجش شباهت رشته‌ای است. ورودی آن دو عبارت رشته‌ای است و خروجی عددی بین صفر تا ۱ است که مشخص‌کننده امتیاز شباهت دو رشته می‌باشد.

۳. استفاده از تکنیک سنجش تشابه معنایی توسط سرویس Wordnet[۱۹] [۳ و ۱۱]. سرویس Wordnet یک پایگاه لغت‌نامه و دانشنامه انگلیسی است که تمامی کلمات و معانی آن را در خود جای داده است. Wordnet لغت‌نامه‌ای است که از یک رابطه هستی‌شناسی[۲۰] برای کلاس‌بندی مفاهیم موجود در دایره لغات استفاده کرده و کلمات انگلیسی اعم از اسم، فعل، صفت و قید را به این مفاهیم منتسب می‌کند. سرویس Wordnet به‌صورت رایگان در اینترنت قابل دسترس است و کتابخانه‌های بسیاری جهت پیاده‌سازی و ارتباط با آن فراهم گردیده است. در روش پیشنهادی برای سنجش شباهت معنایی از این سرویس استفاده می‌شود.

روش پیشنهادی، در طبقه‌بندی سطوح انطباق زیرمجموعه سطح شمای پایگاه داده محسوب شده و در چهار مرحله انطباق را انجام می‌دهد.

۵- راه‌حل پیشنهادی

روش پیشنهادی در چهار مرحله مجزا و به‌صورت سریال مراحل انطباق را انجام می‌دهد:

۵-۱- مرحله پیش پردازش

مرحله اول انطباق که مرحله پیش پردازش نامیده می‌شود کلیه اطلاعات مربوط به اجزای نسخه اولیه، نسخه‌های میانی و نسخه هدف را در یک ماتریس دو بعدی (آرایه دو بعدی) ذخیره می‌نماید. ستون‌های این ماتریس دربرگیرنده اجزای یک طرح از پایگاه داده است. عناصر موجود در سطر این ماتریس هیچ ارتباط ساختاری یا معنایی با یکدیگر ندارند. جدول ۲ عناصر مربوط به سه طرح مختلف یک پایگاه داده را در قالب یک ماتریس نشان می‌دهد:

جدول ۲- جدول عناصر پایگاه‌های داده

Version1	Version2	Version3
Table1	Table3	Table4
Table2	--	Table5

قابل‌ذکر است که در مثال فوق نام‌گذاری عناصر پایگاه هیچ ارتباط معنایی با یکدیگر ندارند و اگر در عمل چنین طرح‌هایی به یک روند انطباق وارد شوند خروجی نهایی نیاز به پالایش خواهد داشت.

پس از آن جهت جلوگیری از مقایسه‌ی نام‌های شبیه به هم در نسخه‌های متفاوت، آن‌ها را اندیس‌گذاری می‌نماییم. اندیس‌گذاری در قالب یک آرایه تک بعدی انجام خواهد شد. اولین اندیس شماره صفر و آخرین اندیس شماره n-1 را می‌پذیرد. n تعداد اسامی غیرمتشابه اجزای نسخ پایگاه‌های داده است:

جدول۳- جدول اندیس گذاری

i[0]= Table1	i[3]=Table4
i[1]= Table2	i[4]= Table5
i[2]=Table3	

۵-۲- مرحله اجرای طرح‌ها

دومین مرحله انطباق، یک Sub workflow بوده که یکی از اصلی‌ترین بخش‌های روند انطباق به شمار می‌رود. این مرحله با استفاده از توابع سنجش رشته‌ای و معنایی دو جدول شباهت رشته‌ای و معنایی را مستقل از هم تولید می‌نماید.

در روش پیشنهادی از دو طرح انطباق به‌صورت موازی استفاده می‌کنیم:

طرح انطباق یک[۲۱]:

- استفاده از تغییرات نسخ میانی در استخراج داده‌های پایگاه دانش
- استفاده از تکنیک سنجش تشابه رشته‌ای توسط الگوریتم Jaro

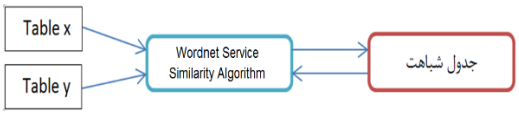
طرح انطباق دو[۲۲]:

- استفاده از تغییرات نسخ میانی در استخراج داده‌های پایگاه دانش
- استفاده از تکنیک سنجش تشابه معنایی توسط سرویس Wordnet

اجرای طرح یک منجر به تولید جدول شباهت رشته‌ای و اجرای طرح دو منجر به تولید جدول شباهت معنایی خواهد شد. این دو جدول فوق خروجی این مرحله محسوب می‌شوند.

۵-۲-۱- ساخت جداول شباهت

جدول شباهت یک ماتریس دو بعدی n*n متقارن است. n تعداد ستون‌های جدول است. ایجاد این جداول در سنجش میزان شباهت هر جزء از پایگاه داده نسبت به اجزاء موجود در پایگاه داده دیگر مورد نیاز است. درصد شباهت‌ها توسط الگوریتم Jaro و سرویس Wordnet به‌صورت مجزا مشخص می‌گردد. شکل ۶ نحوه ساخت جدول شباهت معنایی را نشان می‌دهد:



شکل ۶- روند تکمیل مقادیر جدول شباهت معنایی [۱۳]

در این مرحله دو جدول شباهت وجود دارد. جدول شماره یک جدول شباهت رشته‌ای است. برای تولید این جدول از الگوریتم شباهت Jaro استفاده شده است. جدول شماره دو جدول شباهت‌های معنایی است که برای ساخت آن از Wordnet4 استفاده شده است.

همان‌طور که در جدول شماره ۴ مشاهده می‌شود هم در ستون و هم در سطر، اجزا جهت مقایسه با یکدیگر تکرار می‌شوند. مقادیر سلول‌ها را با عنوان متغیر i1 که عددی بین صفر تا یک است تعریف می‌کنیم. i1 بیانگر میزان شباهت‌های رشته‌ای دو جزء پایگاه داده است.

$$i1=String\_Similarity\ (String\ s1,\ String\ s2) \quad (۱)$$

جدول ۴- جدول شباهت رشته‌ای

	Table1	Table2	Table3	Table4	Table5
Table1	۱				
Table2	۰,۸۹	۱			
Table3	۰,۸۹	۰,۸۹	۱		
Table4	۰,۸۹	۰,۸۹	۰,۸۹	۱	
Table5	۰,۸۹	۰,۸۹	۰,۸۹	۰,۸۹	۱

جدول ۵ با استفاده از سرویس Wordnet تولید شده است. مقادیر سلول‌ها را با عنوان متغیر i2 که می‌تواند مقدارش صفر یا یک باشد تعریف می‌کنیم. i2 بیان‌گر میزان شباهت معنایی دو جزء پایگاه داده است. ۱ معادل True به مفهوم وجود رابطه معنایی و ۰ معادل False به معنای عدم وجود رابطه معنایی است.

جدول ۵- جدول شباهت معنایی

	Table1	Table2	Table3	Table4	Table5
Table1	۱				
Table2	۰	۱			
Table3	۰	۰	۱		
Table4	۰	۰	۰	۱	
Table5	۰	۰	۰	۰	۱

۵-۳- مرحله تولید و ترکیب نتایج

سومین مرحله انطباق ترکیب نتایج تطبیق‌گرها و تولید نتایج ترکیبی است. در این مرحله جدول شباهت نهایی ایجاد شده و براساس آن روابط معنایی جدید تولید می‌گردد [۴]. این روابط معنایی جدید انطباق عنصری در پایگاه اولیه به عنصری در پایگاه هدف را امکان‌پذیر می‌کند.

جدول شباهت نهایی ۶ از ترکیب دو جدول ۴ و ۵ به دست می‌آید. مقدار هر سلول از جدول شماره ۶ با عنوان متغیر i تعریف می‌گردد که برابر با حاصل جمع مقادیر سلول‌های متناظر آن در جداول قبلی است.

$$i = i1 + i2 \quad (۲)$$

بعد از ایجاد جدول شباهت نهایی نیاز به یک شاخص پذیرش با نام S[۲۳] داریم که توسط کاربر (توسعه‌دهنده) مشخص می‌گردد. این شاخص صرفاً مربوط به شباهت رشته‌ای است و کاملاً وابسته به محیط عملیاتی است که پایگاه داده در آن تعریف شده است. به‌عنوان مثال ممکن است در یک محیط، S=۰,۶ باشد به این معنی که میزان شباهت رشته‌ای (i1) بیش از ۰,۶ ملاک پذیرش به‌عنوان یک زوج عناصر منطبق است.

به‌طور کلی حالات زیر برای متغیر شباهت i متصور است:

۱- $i < S$ نشان‌گر عدم شباهت معنایی و رشته‌ای

۲- $S =< i < 1$ نشان‌گر شباهت رشته‌ای

۳- $i = 1$ نشان‌گر شباهت رشته‌ای یا معنایی

۴- $1 < i < 1 + S$ نشان‌گر شباهت معنایی

۵- $1 + S =< i <= 2$ نشان‌گر شباهت معنایی و رشته‌ای

به‌جز حالت اول در بقیه موارد انطباق زوج عنصر متناظر تأیید می‌گردد.

جدول ۶ با استفاده از ترکیب جداول ۴ و ۵ تولید شده است. سلول‌های این جدول می‌توانند مقادیری از صفر تا دو داشته باشند. این مقادیر بیان‌گر میزان شباهت‌های رشته‌ای و معنایی دو جزء پایگاه داده به‌طور هم‌زمان است.

جدول ۶- جدول ترکیبی شباهت

	Table1	Table2	Table3	Table4	Table5
Table1	۱				
Table2	۰,۸۹	۱			
Table3	۰,۸۹	۰,۸۹	۱		
Table4	۰,۸۹	۰,۸۹	۰,۸۹	۱	
Table5	۰,۸۹	۰,۸۹	۰,۸۹	۰,۸۹	۱

پس از ساخت جدول شباهت نهایی نوبت آن است که انطباق اجزا پایگاه مبدأ با پایگاه هدف و با واسطه قرار دادن نسخ میانی به‌صورت زوج-زوج صورت پذیرد. برای رفع نواقص روش‌های قبلی می‌توان از یک مخزن ذخیره داده مثل سیستم‌های مدیریت پایگاه رابطه‌ای و یا XML که قابلیت پرس‌وجو به زبان‌های استاندارد را داشته باشد استفاده کرد.

فرآیند انطباق را به‌صورت بازگشتی از عنصری در پایگاه مبدأ آغاز می‌کنیم. این عنصر را با تمامی عناصر پایگاه‌های بعدی تا پایگاه هدف مقایسه می‌کنیم. اگر مقدار i متناظر هر زوج عنصر مورد مقایسه در جدول ترکیبی شباهت، شرایط پذیرش به‌عنوان یک زوج عنصر منطبق را داشت برای آنها یک رکورد در مخزن اطلاعات ثبت می‌کنیم. قالب رکورد می‌تواند بدین‌صورت تعریف گردد:

(عنصر ۲، امتیاز شباهت، عناصر واسط، عنصر ۱)

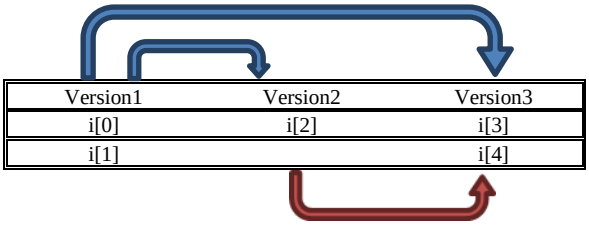
اگر عنصر ۱ در زوج عنصر منطبق یک عنصر از پایگاه‌های واسط باشد در رکورد ثبتی عنصر ۱ به لیست عناصر واسط اضافه شده و عنصر ۲ جایگزین می‌گردد. جدول ۷ این فرآیند را نشان می‌دهد. در این مثال i[0] را با عناصر i[2] و i[3] و i[4] مقایسه می‌نماییم. از آنجایی‌که i[0] با تمامی عناصر ذکر شده منطبق است سه رکورد در مخزن داده ثبت می‌گردد:

(i[0],NULL,0.89,i[2]), (i[0],NULL,0.89,i[3]), (i[0],NULL,0.89,i[4])

سپس رکوردهایی که فیلد عنصر ۲ آنها عنصری در پایگاه هدف نیست واکشی می‌شوند. در مثال فقط رکورد اول این شرط را دارد. با واسطه قرار گرفتن i[2] رکورد یک با رکوردهای زیر جایگزین می‌شود:

(i[0],i[2],1.78,i[3]), (i[0],i[2],1.78,i[4])

جدول ۷- ترتیب مقایسات نسخه های پایگاه داده[۱۳]



این رویه به‌صورت بازگشتی تا واسط قرار گرفتن n-1 امین پایگاه ادامه پیدا می‌کند.

برای اطلاع از نتیجه انطباق می‌توان با پرس‌وجوهای مختلف گزارشات متفاوتی را از مخزن داده دریافت کرد.

با واکشی رکوردهایی که عنصر ۲ آنها جزئی از پایگاه هدف است لیست زوج عناصر منطبق شده استخراج می‌گردد.

۵-۴- مرحله پالایش نتایج

در این مرحله گزارش کاملی از انطباق‌های یکتا، موازی و چندگانه به کاربر اعلام می‌گردد تا براساس نوع محیط عملیاتی در خصوص حذف یا انتخاب آنها تصمیم‌گیری شود.

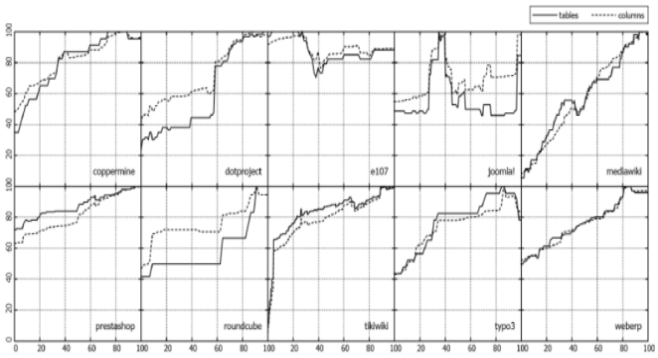
۶- ارزیابی

پیاده‌سازی طرح پیشنهادی با استفاده از زبان برنامه‌نویسی JAVA و بسته توسعه JDK 8u121 و در بستر NetBeans IDE 8.2 انجام گرفته است. علت انتخاب این بستر برنامه‌نویسی امکان اجرای آن بر روی هر سخت‌افزار و سیستم عاملی است که برای آن یک ماشین مجازی جاوا وجود داشته باشد. سخت‌افزار مورد استفاده جهت پیاده‌سازی و تست، کامپیوتری با پردازنده Core(TM) i7 1.80 GHz و حافظه اصلی ۴ گیگا بایت و سیستم عامل ویندوز ۱۰ بوده است. در پیاده‌سازی سعی گردیده از کتابخانه غنی JDK استفاده گردد.

به‌دلیل این که در عملیات انطباق مقایسات آیتم‌ها مکرراً تکرار می‌گردند تمامی ساختار و ساختمان داده‌ها ابتدا به حافظه اصلی بارگذاری گردیده و سپس عملیات مقایسه‌ای انجام می‌گیرد. این عمل از آن حیث قابل اهمیت است که باعث بالا رفتن چشمگیر سرعت انطباق می‌شود.

جهت ارزیابی طرح پیشنهادی در این تحقیق، آن را با نتایج به‌دست آمده از برترین تطبیق‌گرهای شرکت‌کننده در مسابقات OAEI[۲۴] در سال‌های ۲۰۰۶ تا ۲۰۱۵ مورد مقایسه قرار خواهیم داد تا دقت انطباق آن در مقایسه با روش‌های مطرح مورد بررسی قرار گیرد. ابزار و محیط تست همانند ابزارها و متدهای سنجش در روش ۳-۴ است [۱۳] که به‌عنوان یکی از ستون‌های ترکیب در روش پیشنهادی انتخاب شده است تا بتوان در شرایطی برابر نتایج حاصل از ارزیابی را مقایسه نمود. همچنین با توجه به ضرورت و اهمیت ارزیابی برای اثبات برتری و کارآیی این روش در انطباق، آن را در چندین نوع Dataset مورد ارزیابی قرار می‌دهیم.

معیار سنجش عملکرد روش پیشنهادی پارامترهای Precision، Recall[۲۵] و F-measure هستند [۳]، این پارامترها معیار سنجش عملکرد تطبیق‌گرها در مسابقات OAEI بوده‌اند [۱۰].



شکل ۷- روند سیر تکاملی جداول و ستون‌های پایگاه داده‌ها در برخی پروژه‌های مطالعه شده. محور افقی روند پردازش و پیشرفت پایگاه داده و محور عمودی تعداد جدول و ستون اضافه یا حذف شده در پایگاه داده پروژه است [۲]

برای انجام این آزمایش‌ها از پنج سری پایگاه داده‌های واقعی متن‌باز: RoundCube, DotProject, phpMyAdmin, Joomla، استفاده شده است.

دلیل اینکه از مجموعه داده‌های استاندارد OAEI استفاده نشده است، عدم وجود سری نسخه‌ها برای این داده‌ها است. [۱۳] میزان تغییرات بوجود آمده در نسخه‌ها در پایگاه داده‌های مذکور توسط Qiu و همکاران در سال ۲۰۱۳ انجام گردیده است [۷] که در شکل ۷ قابل مشاهده است.

قابل‌ذکر است که تعداد سیستم‌های شرکت‌کننده در مسابقات OAEI2016 نسبت به سال ۲۰۱۵ و ۲۰۱۳ کمتر بوده اما نسبت به سال ۲۰۱۴ افزایش داشته است. در این مسابقات ترکیبی از سیستم‌های جدید و قبلی حضور داشته‌اند. سیستم‌هایی که در سال‌های قبل شرکت کرده بودند نیز امتیاز مشابهی را کسب کرده‌اند که نشان‌دهنده این است که توسعه قابل توجهی صورت نپذیرفته است [۱۲].

زمان اجرای الگوریتم از مرحله بارگذاری ساختار پایگاه داده در حافظه اصلی تا مرحله نمایش نتایج "6.01 ثانیه می‌باشد. جدول ۸ گزارشی از سنجش عملکرد چند تطبیق‌گر معروف بین سال‌های ۲۰۰۵ تا ۲۰۱۵ را نشان می‌دهد تا امکان مقایسه روش پیشنهادی با این روش‌ها فراهم شود.

جدول ۸- مقایسه روش‌های مطرح با میانگین

سال ارائه	Recall	F-measure	Precision	Matchers
۲۰۱۶	۰,۹۱	۰,۹۱	۰,۹۳	روش ترکیبی - معنایی
۲۰۱۴	۰,۷۸	۰,۸۰	۰,۸۲	AML
۲۰۱۳	۰,۷۵	۰,۷۳	۰,۷۲	AML 2013
۲۰۱۵	۰,۷۵	۰,۶۹	۰,۶۴	بدون نسخه واسط
۲۰۱۰	۰,۶۵	۰,۶۳	۰,۶۱	ASMOV 2010
۲۰۰۹	۰,۶۵	۰,۶۲	۰,۶	ASMOV 2009
۲۰۰۶	۰,۲۸	۰,۳۰	۰,۳۲	COMA++
۲۰۰۵	۰,۲۶	۰,۲۸	۰,۳۱	COMA

۷- نتیجه‌گیری

با توجه به مزایا و ویژگی‌های ارائه شده در روش پیشنهادی و اطلاعات مفیدی که استفاده توام از نسخه‌های میانی پایگاه داده‌ها، سنجش تشابه رشته‌ای و همچنین سنجش تشابه معنایی به ما می‌دهد دقت استنتاج‌های به‌عمل آمده در تولید روابط معنایی جدید بسیار افزایش می‌یابد و باعث انطباقی دقیق‌تر و صحیح‌تر نسبت به روش‌های انطباق دیگر خواهد شد. از طرفی بدلیل این که مراحل چهارگانه انطباق به‌صورت سریال باید انجام گیرد انتظار داریم در مقایسه با روش‌های انطباق نمونه محور زمان بیشتری سپری شود.

همانطور که در جدول ۸ قابل مشاهده است روش پیشنهادی به مراتب از روش‌های دیگر و همچنین روش ۳-۴ بهتر است. روش پیشنهادی در حقیقت کلیه تکنیک‌های روش‌های اطلاعات محور (coma++) و روش ترکیبی استفاده از نسخ میانی را به ارث برده و با اضافه نمودن تکنیک سنجش شباهت معنایی در تولید جدول شباهت نهایی دقت انطباق را به مراتب افزایش داده است.

براساس آزمایشات به‌عمل آمده هنگامی که یک نسخه از سری نسخه‌های پایگاه داده دارای تغییر ناگهانی شود و یا پایگاه دانشی که در توابع سنجش میزان شباهت معنایی از آن استفاده می‌شود با واقعیات محیط عملیاتی منطبق نباشد، روش پیشنهادی نمی‌تواند نتایج صحیحی را ارائه نماید.

مراجع

[1] H. Seifzadeh, H. Abolhassani, and M. S. Moshkenani, "A survey of dynamic software updating," *Journal of Software: Evolution and Process*, vol. 25, no. 5, pp. 535-568, 2013.

[2] E. Rahm, "Towards large-scale schema and ontology matching," *Schema matching and mapping*, pp. 3-27: Springer, 2011.

[3] J. Euzenat, and P. Shvaiko, "Ontology matching," *Springer Science & Business Media*, 2013.

[4] P. Shvaiko, and J. Euzenat, "A survey of schema-based matching approaches," *Journal on data semantics IV. Springer Berlin Heidelberg*, pp. 146-171, 2005.

[5] A. Islam, and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 2, no. 2, pp. 10, 2008.

[6] Y. R. Jean-Mary, E. P. Shironoshita, and M. R. Kabuka, "Ontology matching with semantic verification," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, pp. 235-251, 2009.

[7] D. Qiu, B. Li, and Z. Su, "An empirical analysis of the co-evolution of schema and code in database applications," *In Proceedings of the 9th Joint Meeting on Foundations of Software Engineering*, pp. 125-135, 2013.

[8] H. H. Do, and E. Rahm, "Matching large schemas: Approaches and evaluation," *Information Systems*, vol. 32, no. 6, pp. 857-885, 2007.

[9] G. Kappel, H. Kargl, G. Kramler, A. Schauerhuber, M. Seidl, M. Strommer, and M. Wimmer, "Matching Metamodels with Semantic Systems-An Experience Report," *In BTW workshops*, pp. 38-52, 2007.

[10] Z. Dragisic, K. Eckert, J. Euzenat, D. Faria, A. Ferrara, R. Granada, V. Ivanova, E. Jiménez-Ruiz, A. O. Kempf, and P. Lambrix, "Results of the ontology alignment evaluation initiative 2014," *In Proceedings of the 9th International Conference on Ontology Matching-Volume*, pp. 61-104, 2014.

[11] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains," *Expert Systems with Applications*, vol. 42, no. 4, pp. 2264-2275, 2015.

[12] M. Achichi, M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, and E. Jiménez-Ruiz, "Results of the Ontology Alignment Evaluation Initiative 2016," *In 11th ISWC workshop on ontology matching (OM)*, pp. 73-129, 2016.

[۱۳] ع. بهرامی بوانی، استفاده از تاریخچه تغییرات نسخه‌ها برای تطبیق و بروزرسانی سیستم‌های پایگاه داده، پایان‌نامهٔ کارشناسی ارشد نرم‌افزار، گروه کامپیوتر، دانشگاه آزاد اسلامی واحد اصفهان، ۱۳۹۴.

محمد ظهیری دستجردی فارغ‌التحصیل از دانشگاه آزاداسلامی نجف‌آباد در مقطع کارشناسی و دانشگاه گیلان در مقطع کارشناسی‌ارشد در رشته مهندسی کامپیوتر گرایش نرم‌افزار. زمینه‌های مورد علاقه سیستم‌های مدیریت پایگاه داده، توسعه نرم‌افزارهای تحت وب، توسعه نرم‌افزارهای تحت سیستم عامل اندروید، سیستم‌های توزیع شده و رایانش ابری هستند.
آدرس پست‌الکترونیکی ایشان عبارت است از:
m.zahiri.d@gmail.com

حمیدرضا احمدی‌فر فارغ‌التحصیل از دانشگاه شهید بهشتی در مقاطع کارشناسی و دکتری و دانشگاه صنعتی امیرکبیر در مقطع کارشناسی‌ارشد در رشته مهندسی کامپیوتر با گرایش معماری سیستم‌های کامپیوتری. از سال ۱۳۸۲ عضو هیات علمی گروه مهندسی کامپیوتر دانشگاه گیلان بوده و زمینه‌های مورد علاقه حساب کامپیوتری، سیستم‌های توزیع شده، رایانش ابری و بکارگیری شبکه‌های عصبی در حل مسائل مختلف هستند.
آدرس پست‌الکترونیکی ایشان عبارت است از:
ahmadifar@guilan.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۱۲/۲۱
تاریخ اصلاح: ۱۳۹۶/۱۲/۲۱
تاریخ قبول شدن: ۱۳۹۷/۰۲/۰۵
نویسنده مرتبط: دکتر حمیدرضا احمدی‌فر، دانشکده فنی و مهندسی، دانشگاه گیلان، رشت، ایران.

---

[1]Update
[2]Match
[3]Database Schema
[4]String Similarity
[5]Semantic Similarity
[6]Matchers
[7]Pre-Processing
[8]Name Similarities
[9]Sub Workflow
[10]Element Level
[11]Structure Level
[12]Instance-Based Matching
[13]Ontology Alignment Evaluation Initiative - (http://oaei.ontologymatching.org)
[14]Loading Module
[15]Fact
[16]Ruls
[17]Taxonomy-Based Techniques
[18]Prolog
[19]WordNet Is a Lexical Database for the English Language
[20]Ontology
[21]Matcher 1
[22]Matcher 2
[23]Similarity Degree
[24]Ontology Alignment Evaluation Initiative - (http://oaei.ontologymatching.org)
[25]F-Measure Combines Recall and Precision, Two Standard Measures to Evaluate the

The CSI Journal on Computing Science and Information Technology

The CSI Journal on
Computing Science and
Information Technology
Vol. 16, No. 1, 2018

Updating and Semantic Matching of Different Versions of Database Systems

Mohammad Zahiri Dastjerdi **Hamidreza Ahmadifar**

Department of Electrical Engineering, University of Guilan, Rasht, Iran

ABSTRACT

In this paper, according to the articles and researches presented until 2016, the most widely used methods to match the components of early versions and the purpose of a database, along with their advantages and disadvantages are discussed. Next, matching bugs are especially examined based on string similarity measure using changes in intermediate versions, and finally, a solution is proposed in which information for component changes in intermediate versions can be used, and also semantic similarity measure can be used in addition to string similarity measure so that the probability of error is significantly reduced. To do so, two separate and independent similarity tables are used that are produced based on string and semantic similarities and a final similarity table that combines the elements of two previous tables. Finally, experimental results based on comparison with the other methods for database matching show that our solution has better accuracy and correction.

Keywords: Database, Matcher, Matching, Semantic Similarity.