

4

Alinhamento de esquemas conceituais

4.1

Introdução

Catálogos são bancos de dados descritos com esquemas conceituais simples que utilizam um esquema de classificação acoplado para classificar seus objetos em categorias pré-definidas. A abordagem para alinhamento de esquemas conceituais simples proposta neste capítulo utiliza a *técnica de sondagem de consultas* introduzida em (Wang et al., 2004). Esta técnica consiste em submeter consultas às fontes de dados utilizando valores de atributos de um conjunto de *instâncias de referência*. De posse dos resultados destas consultas, as co-ocorrências dos valores dos atributos utilizados nas consultas nos conjuntos de instâncias retornados pelas fontes são contadas. Estes valores co-ocorrentes computam evidências para os mapeamentos dos atributos dos esquemas conceituais a serem integrados.

A técnica de sondagem de consultas será utilizada de duas formas distintas. Na primeira forma, descrita na seção 4.2, o esquema global e um conjunto de instâncias de referência são previamente definidos. A técnica de sondagem de consultas é então aplicada para criar mapeamentos dos esquemas das fontes a serem integradas para o esquema global.

A segunda forma, descrita na seção 4.3, segue um enfoque adaptativo. As instâncias de referência utilizadas são as instâncias retornadas a partir de pares de consultas realizadas pelo usuário durante uma seção e direcionadas a duas fontes de dados distintas. Os valores dos atributos das instâncias retornadas de uma fonte são comparados com as instâncias retornadas de outra fonte, gerando mapeamentos entre os esquemas das fontes a serem integradas.

4.2

Abordagem *a priori*

Nesta seção será introduzida uma abordagem *a priori* para alinhamento de esquemas conceituais utilizando uma coleção de instâncias de referência globais

e um esquema global, previamente definidos, como entrada do processo de alinhamento.

A abordagem é baseada no processo ilustrado na Figura 17.

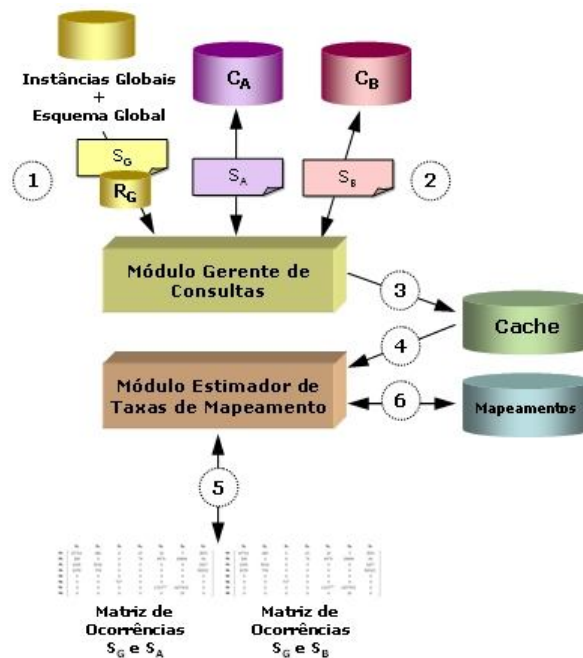


Figura 17 – O processo de alinhamento *a priori* de esquemas conceituais.

O processo inicia com uma coleção de instâncias globais R_G , um esquema global S_G e os metadados contendo as descrições dos esquemas conceituais S_A e S_B , das fontes C_A e C_B a serem integradas (passo 1 da Figura 17).

O *Módulo Gerente de Consultas* (MGC) é responsável por, a partir de cada instância do conjunto R_G , formular consultas para C_A e C_B utilizando os valores das instâncias de referência. O MGC então submete as consultas às fontes cadastradas (passo 2).

Os resultados das consultas, R_A e R_B , são então armazenados em um cache local para posterior análise (passo 3) pelo *Módulo Estimador de Taxas de Mapeamento* (METM).

O METM é responsável por acessar o cache, recuperar os resultados armazenados, R_A e R_B , e analisar os valores recorrentes (passo 4).

O METM analisa as instâncias contando as co-ocorrências dos valores dos atributos das instâncias globais nos valores dos atributos das instâncias retornadas de cada fonte. Assim, o METM gera matrizes de ocorrências entre S_G e S_A (matriz A) e entre S_G e S_B (matriz B) (passo 5).

Em uma matriz de ocorrências A, as colunas representam os atributos do esquema global $S_G = \{n_1, n_2, n_3, \dots, n_j\}$ e as linhas representam os atributos do

esquema a ser alinhado, $S_A = \{m_1, m_2, m_3, \dots, m_i\}$. Cada elemento a_{ij} da matriz A representa o número de co-ocorrências dos valores do atributo n_j das instâncias globais como valor do atributo m_i do esquema S_A .

Para descobrir os mapeamentos, as matrizes de ocorrências A e B devem ser normalizadas, utilizando as fórmulas (3) e (4), respectivamente. Após a normalização, são obtidas as matrizes de mapeamentos E e F . O processo de normalização adotado divide os valores de co-ocorrência (cada elemento da matriz A ou B) pelo número total de instâncias retornadas no conjunto de resultado correspondente (R_A ou R_B).

$$e_{ij} = \frac{a_{ij}}{|R_A|} \quad \text{e} \quad f_{ij} = \frac{b_{ij}}{|R_B|} \quad (3) \text{ e } (4)$$

Onde:

e_{ij} elemento da matriz E da i -ésima linha e j -ésima coluna.

a_{ij} elemento da matriz A da i -ésima linha e j -ésima coluna.

$|R_A|$ número de instâncias do conjunto R_A .

f_{ij} elemento da matriz F da i -ésima linha e j -ésima coluna.

b_{ij} elemento da matriz B da i -ésima linha e j -ésima coluna.

$|R_B|$ número de instâncias do conjunto R_B .

Com base na validação descrita na seção 4.4.2, foi definido que um par de atributos (n_j, m_i) alinham sse o elemento da matriz E ou F , e_{ij} ou f_{ij} , for maior ou igual ao limiar de 0.2 (ou seja, 20%). Este valor foi definido com base nos mapeamentos corretos obtidos através do experimento descritos na seção 4.4.2 e em (Brauner et al., 2007b). Os mapeamentos são então armazenadas em um banco de dados local de mapeamentos (passo 6).

A abordagem descrita nesta seção pode ser utilizada em um mediador de consultas a bancos de dados disponíveis via Web Services. Em (Brauner et al., 2007b) é apresentado um experimento utilizando esta abordagem para alinhamento de esquemas de exportação de bancos de dados geográficos disponíveis via Web services. A abordagem de alinhamento é executada *a priori* e os mapeamentos descobertos são utilizados pelo mediador para mediar

consultas aos Web Services utilizando, como base para as consultas, o esquema global previamente definido. Este experimento é detalhado na seção 4.4.2 desta tese.

4.3 Abordagem adaptativa

Nesta seção será introduzida uma abordagem adaptativa para alinhamento de esquemas conceituais que diferente da anterior por não utilizar uma coleção de instâncias globais e um esquema global. Esta abordagem segue um enfoque adaptativo. As instâncias de referência utilizadas são as instâncias retornadas a partir de pares de consultas realizadas pelo usuário durante uma sessão e direcionadas a duas fontes de dados distintas. Os valores dos atributos das instâncias retornadas de uma fonte são comparados com as instâncias retornadas de outra fonte, gerando mapeamentos entre os esquemas das fontes a serem integradas.

A abordagem de alinhamento de esquemas conceituais é baseada no processo ilustrado na Figura 18.

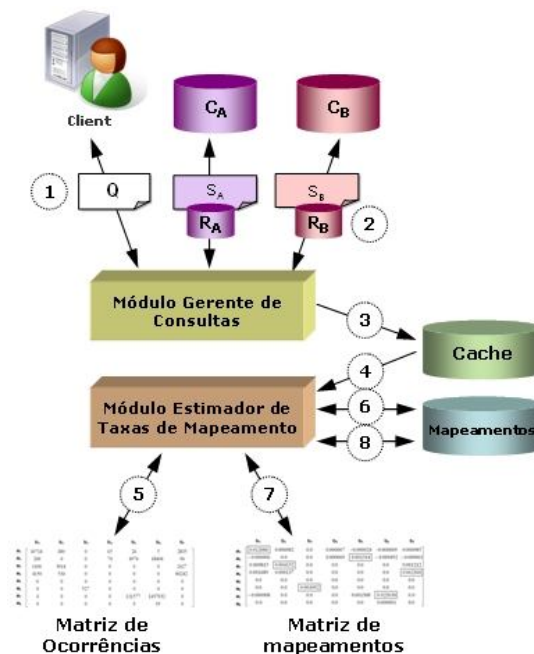


Figura 18 – O processo de alinhamento *adaptativo* de esquemas conceituais.

Suponha um mediador fornecendo acesso a fontes de dados C_A e C_B com esquemas conceituais S_A e S_B , respectivamente. O processo inicia com uma consulta por palavra-chave Q do usuário (ou cliente) submetida ao mediador

(passo 1 da Figura 18). O mediador deve ter também os metadados contendo as descrições dos esquemas conceituais S_A e S_B (passo 1 da Figura 18).

O *Módulo Gerente de Consultas* (MGC) é responsável por criar as consultas, Q_A e Q_B , para cada fonte e encaminhá-las às fontes cadastradas (passo 2). A consulta para cada fonte é criada utilizando as informações de cada fonte previamente definidas no momento de cadastramento da fonte no mediador. Quando registradas, as fontes informam o serviço de consulta por palavra-chave disponível, o qual utiliza um ou mais campos de consultas.

Os resultados das consultas, os conjuntos R_A e R_B , são então devolvidos aos usuários e, simultaneamente, armazenados em um cache local para posterior análise (passo 3).

O *Módulo Estimador de Taxas de Mapeamento* (METM) é responsável por acessar o cache, recuperar os conjuntos de instâncias armazenadas, R_A e R_B , e analisar os valores co-ocorrentes (passo 4).

O METM analisa as instâncias contando as co-ocorrências dos valores dos atributos das instâncias retornadas de uma fonte de dados nos valores dos atributos das instâncias retornadas da outra fonte. Assim, o METM gera uma matriz de ocorrências P contendo os valores co-ocorrentes entre S_A e S_B (passo 5). Vale salientar que os valores co-ocorrentes são computados utilizando uma junção natural, implementada através de uma consulta SQL. Em virtude disso, o valor de co-ocorrência de um elemento da matriz P será no máximo igual à maior cardinalidade entre R_A e R_B .

Em uma matriz de ocorrências P , as colunas representam os atributos do esquema $S_A = \{n_1, n_2, n_3, \dots, n_j\}$ e as linhas representam os atributos do esquema $S_B = \{m_1, m_2, m_3, \dots, m_i\}$. Cada elemento p_{ij} da matriz P representa o número de co-ocorrências dos valores do atributo m_i do esquema S_B como valor do atributo n_j do esquema S_A .

Para viabilizar o acúmulo de evidências no mediador, ou seja, para que o processo de alinhamento funcione de forma adaptativa, o METM armazena algumas informações em um banco de dados local de mapeamentos (passo 6), tais como: a matriz de ocorrências e o número total de instâncias retornadas em cada um dos conjuntos R_A e R_B .

No passo 7, o METM recupera a matriz de ocorrências e o número total de instâncias armazenados no banco de dados de mapeamentos para calcular a matriz de mapeamentos utilizando a fórmula (5).

$$g_{ij} = \frac{\Delta(p_{ij}) + \alpha(n(p_{ij}) + \psi)}{\Delta(N_R) + \alpha(n(N_R) + 1)} \quad (5)$$

Onde:

- g_{ij} elemento da i -ésima linha e j -ésima coluna da matriz de mapeamentos G .
- p_{ij} elemento da i -ésima linha e j -ésima coluna da matriz de ocorrências P .
- $\Delta(p_{ij})$ o valor de p_{ij} referente às evidências antigas, ou seja, valor de p_{ij} da matriz P recuperada do cache.
- $n(p_{ij})$ valor de p_{ij} referente às novas evidências levantadas a partir dos resultados das últimas consultas realizadas.
- N_R maior número de instância entre os conjuntos de resultados, dado por $\max(|R_A|, |R_B|)$.
- $\Delta(N_R)$ valor de N_R referente às evidências antigas, ou seja, dos valores de R_A e R_B armazenados em cache.
- $n(N_R)$ valor referente às novas evidências de N_R , levantadas a partir dos resultados das últimas consultas realizadas.
- $|R_A|$ número de instâncias do conjunto R_A .
- $|R_B|$ número de instâncias do conjunto R_B .
- α coeficiente que assume um dos valores do conjunto $\{0.01, 0.1, 0, 1, 10, 100\}$. Este coeficiente pode ser calibrado durante um processo de validação. Nesta tese assumimos o mesmo valor calibrado na seção 3.4.2 por se tratar de uma adaptação da mesma fórmula. Neste contexto, 100 indica dar maior importância às evidências antigas enquanto 0.01 indica dar maior importância às novas evidências.
- $\psi = \frac{1}{N_R}$ coeficiente de suavização dado pelo inverso do tamanho do maior conjunto de resultados. OBS: o tamanho do conjunto de resultados é dado pelo número de registros que ele possui.

Os mapeamentos são então descobertos a partir da matriz de mapeamentos gerada. Estes mapeamentos são então armazenados no banco de dados local de mapeamentos (passo 8). Os atributos com taxas de mapeamento acima de 0.2 alinham. Este limiar de 0.2 foi adotado a partir de uma breve validação realizada nos resultados obtidos nos testes descritos na seção 4.4.2.

A abordagem descrita nesta seção pode ser utilizada em um mediador para processar consultas a fontes de dados num mesmo domínio, mas com esquemas conceituais heterogêneos. Em (Brauner et al., 2008) é apresentado um experimento utilizando uma versão preliminar desta abordagem para alinhamento de esquemas de exportação de bancos de dados disponíveis via Web Services. A abordagem de alinhamento é executada de forma adaptativa e os mapeamentos descobertos são utilizados pelo mediador para mediar consultas aos Web Services. Foram realizados dois experimentos: um no domínio de dados geográficos e outro no domínio de livros.

O processo ilustrado pela Figura 18 é aplicado a um mediador de consultas a fontes de dados com esquemas heterogêneos. A Figura 19 mostra a arquitetura proposta para um mediador utilizando esta abordagem.

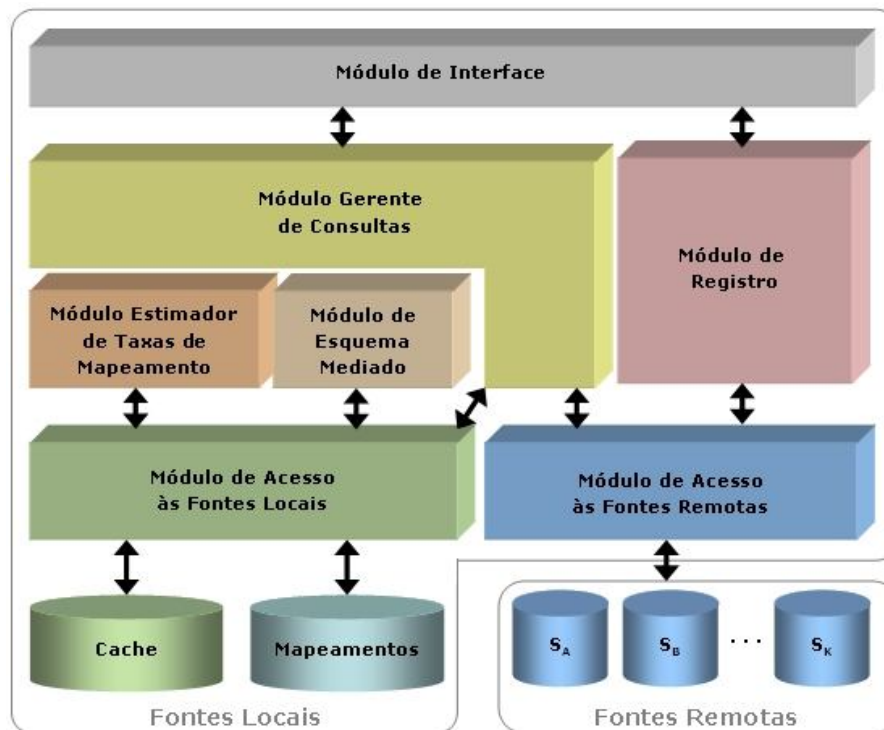


Figura 19 – Arquitetura proposta para um mediador utilizando a abordagem *adaptativa* para alinhamento de esquemas.

O *Módulo de Interface* (MI) é responsável pela comunicação entre os usuários (clientes) e o mediador. O MI recebe as consultas dos usuários e retorna seus resultados. Ele se comunica com o *Módulo de Registro* (MR) para cadastrar, editar e excluir fontes de dados. Por sua vez, o MR comunica-se com o *Módulo de Acesso às Fontes Remotas* (MAFR) para recuperar as informações

de metadados das fontes. O MI comunica-se também com o *Módulo Gerente de Consultas* (MGC).

O MGC é responsável por decompor as consultas dos usuários em subconsultas, reescrevendo-as no formato padrão das fontes de dados cadastradas, e submetê-las. O MGC comunica-se com o *Módulo de Acesso às Fontes Locais* (MAFL) para acessar as fontes de dados locais de cache e de mapeamentos. Durante o processo de decomposição das consultas, o MGC comunica-se com o banco de dados de mapeamentos para recuperar os mapeamentos existentes e formular as consultas no formato das fontes de dados. Durante o processo de submissão e retorno das consultas, o MGC comunica-se com o MAFL para submeter as consultas às fontes e receber os resultados, e então retornar os resultados das consultas ao cliente e armazenar os resultados no cache local do mediador através do MAFL.

O *Módulo Estimador de Taxas de Mapeamento* (METM) é um módulo autônomo, responsável por acessar as fontes locais de cache e mapeamentos para computar os alinhamentos para os elementos dos esquemas conceituais, conforme explicado na seção anterior.

Com base nos mapeamentos descobertos, o *Módulo de Esquema Mediado* (MEM) cria um esquema mediado para que os resultados possam ser apresentados ao usuário (cliente) num formato único. Nesta tese, não será explorada a geração de esquema mediado, portanto assume-se que o MEM usa uma heurística simples para determinar os atributos do esquema mediado: o esquema mediado conterá todos os elementos dos esquemas das fontes cadastradas.

4.4

Validação e testes

Esta seção descreve experimentos para validação e teste das abordagens apresentadas nas seções 4.4.2 e 4.4.3, utilizando as fontes descritas na seção 4.4.1.

4.4.1

Introdução

Nestes experimentos, utilizamos dois *gazetteers* disponíveis para consultas através da Web para avaliar as abordagens de alinhamento de

esquemas: o Geonames e o *Alexandria Digital Library Gazetteer* (ADL Gazetteer).

O GeoNames (GEONAMES, 2008) é um banco de dados de geográficos disponível gratuitamente para download ou consulta via Web Services. O GeoNames integra dados de várias fontes, tais como: o GEOnet Names Server, o *U.S. Geological Survey Geographic Names Information System*, a GeoBase – uma base de dados geográficos do Canadá mantida pelo governo canadense, entre outros. A Figura 20 apresenta um fragmento de um XML retornado por uma operação dos serviços oferecidos pelo GeoNames.

O *Alexandria Digital Library Gazetteer* (ADL Gazetteer) já foi descrito na seção 3.4.1. O acesso ao ADL Gazetteer é disponibilizado através do *ADL Gazetteer Protocol* (Janée & Hill, 2004), um protocolo baseado em requisições HTTP e XML. A Figura 21 apresenta um fragmento de um XML retornado por uma das operações do serviço oferecidos pelo ADL Gazetteer.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <geonames style="FULL">
  <totalResultsCount>1</totalResultsCount>
  - <geoname>
    <name>Amazon River</name>
    <lat>-0.1666667</lat>
    <lng>-49.0</lng>
    <geonameId>3407729</geonameId>
    <countryCode>BR</countryCode>
    <countryName>Brazil</countryName>
    <fcl>H</fcl>
    <fcode>STM</fcode>
    <fclName>stream, lake, ...</fclName>
    <fcodeName>stream</fcodeName>
    <population />
    <alternateNames>Orellana,Rio Amazonas,Rio Marañon,Rio Solimoes,Rio
      Solimões,Rio el Amazonas,Rio Amazonas,Rio Marañón,Rio el
      Amazonas,Solimoes River,Solimoens</alternateNames>
    <elevation />
    <adminCode1>00</adminCode1>
    <adminName1 />
    <adminCode2 />
    <adminName2 />
    <timezone dstOffset="-3.0" gmtoffset="-3.0">America/Belem</timezone>
  </geoname>
</geonames>
```

Figura 20 – Fragmento do XML de retorno do serviço de consulta do GeoNames.

```

<?xml version="1.0" encoding="UTF-8" ?>
- <gazetteer-service xmlns="http://www.alexandria.ucsb.edu/gazetteer"
  xmlns:gml="http://www.opengis.net/gml" xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.alexandria.ucsb.edu/gazetteer
    http://www.alexandria.ucsb.edu/gazetteer/protocol/gazetteer-service.xsd" version="1.2">
- <query-response>
- <standard-reports>
- <gazetteer-standard-report>
  <identifier>adlgaz-1-1410143-3a</identifier>
  <place-status>current</place-status>
  <display-name>Amazon River - Brazil</display-name>
- <names>
  <name primary="true" status="current">Amazon River</name>
  <name primary="false" status="current">Solimoens</name>
  <name primary="false" status="current">Salimoes River</name>
  <name primary="false" status="current">Orellana</name>
  <name primary="false" status="current">Maranon, Rio</name>
  <name primary="false" status="current">Amazonas, Rio</name>
  <name primary="false" status="current">Amazonas, Rio el</name>
  <name primary="false" status="current">Solimoens, Rio</name>
</names>
- <bounding-box>
- <gml:coord>
  <gml:X>-49.0</gml:X>
  <gml:Y>-0.1667</gml:Y>
</gml:coord>
- <gml:coord>
  <gml:X>-49.0</gml:X>
  <gml:Y>-0.1667</gml:Y>
</gml:coord>
</bounding-box>
- <footprints>
- <footprint primary="true">
  - <gml:Point>
    - <gml:coord>
      <gml:X>-49.0</gml:X>
      <gml:Y>-0.1667</gml:Y>
    </gml:coord>
    </gml:Point>
  </footprint>
</footprints>
- <classes>
  <class thesaurus="ADL Feature Type Thesaurus" primary="true">streams</class>
  <class thesaurus="NIMA Feature Designation" primary="false">STM (stream)</class>
</classes>
- <relationships>
  <relationship relation="part of" target-name="UTM grid GE28" />
  <relationship relation="part of" target-name="JOG Sheet Number SA22-0" />
  <relationship relation="part of" target-name="Brazil" target-identifier="adlgaz-1-19-19" />
</relationships>
</gazetteer-standard-report>
</standard-reports>
</query-response>
</gazetteer-service>

```

Figura 21 – Fragmento do XML de retorno do serviço de consulta do ADL Gazetteer.

4.4.2

Validação e teste da abordagem *a priori*

Como detalhado na seção 4.2, a abordagem *a priori* para alinhamento de esquemas usa um conjunto de instâncias de referência previamente selecionadas e classificadas segundo um esquema conceitual global (ou

simplesmente *esquema global*) também previamente definido. Nesta abordagem, são criados mapeamentos dos esquemas das fontes a serem integradas com o esquema global.

Os elementos dos esquemas utilizados neste experimento são detalhados a seguir. A Tabela 10 e a Tabela 11 apresentam os esquemas de exportação da fonte Geonames e da fonte ADL Gazetteer, respectivamente, utilizadas neste experimento.

Tabela 10 – Esquema de exportação da fonte GeoNames.

Attribute name	Description	Data type
geonameId	The entry identifier	String
name	The entry primary name	String
alternateNames	Comprises the set of alternative names	String
countryCode	The entry country code (ISO-3166 2-letter code)	String
countryName	The entry country name	String
population	The population of the instance	Number
lat	The entry latitude	Number
lng	The entry longitude	Number
fcl	The feature type super class code	String
fclName	The feature type super class name	String
fcode	The feature type classification code	String
fcodeName	The feature type classification name	String
elevation	The entry elevation, in meters	Number
admCode1	Code for first administrative division	String
admName1	Name for first administrative division	String
admCode2	Code for second administrative division	String
admName2	Name for second administrative division	String
timezone	Timezone description	String

Tabela 11 – Esquema de exportação da fonte ADL Gazetteer.

Attribute name	Description	Data type
identifier	The entry identifier	String
placeStatus	The entry place-status (current or former)	String
name	The entry primary name	String
displayName	The entry primary name as it is displayed	String
footprintX	The entry longitude	Number
footprintY	The entry latitude	Number
class	The entry class	String
thesaurus	The thesaurus of the entry class	String
names	Comprises the set of alternative names	String
relationships	The entry "partOf" relationships	String

Para condução deste experimento, foi criado um esquema conceitual global capturando as características essenciais para *gazetteers* seguindo as normas definidas pela ISO (ISO/TC211 ISO19112, 2003). A Figura 22 mostra o modelo Entidade-Relacionamento (E-R) do esquema conceitual global criado. Em detalhes, o esquema global contém duas classes principais, *GeoInstance* e *GeoType*, que são baseadas, respectivamente, nas classes recomendadas pela ISO, *SI_LocationInstance* e *SI_Location Type*. A Tabela 12 e a Tabela 13 mostram, respectivamente, os atributos das classes *GeoInstance* e *GeoType*.

Conforme o processo descrito na seção 4.2, as instâncias de referência representam os dados que serão submetidos como consultas às fontes a serem integradas. Por isso, elas foram selecionadas com base em lugares geográficos populares. Assim, há mais chance de elas ocorrerem nas fontes a serem integradas. Para selecionar as instâncias, primeiramente, foram listados 36 nomes de locais populares. Estes foram submetidos como consulta ao *gazetteer* GeoNames. Como esperado, para cada nome consultado foram retornados diversos objetos geográficos como resultado contendo a ocorrência do nome consultado. Estes resultados foram manualmente verificados de forma a encontrar os objetos reais que representam as 36 instâncias de referência, os demais foram descartados. A Tabela 14 mostra um fragmento do conjunto de instâncias de referência resultante.

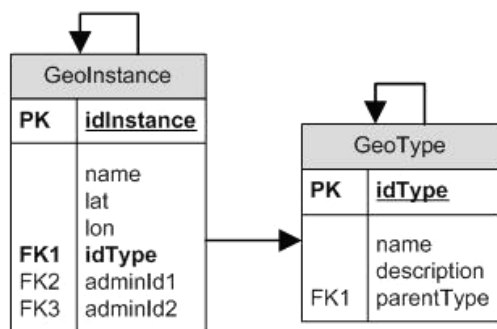


Figura 22 – Modelo E-R do esquema global utilizado.

De posse do conjunto de instâncias de referência e do esquema global, partimos para os experimentos seguindo o processo descrito na seção 4.1. Para as 36 instâncias de referência submetidas como consultas aos *gazetteers* introduzidos na seção 4.4, foram obtidas 459 instâncias do ADL Gazetteer e 703 do GeoNames.

Tabela 12 – Atributos da classe *GeoInstance* do esquema global.

Attribute name	Description	Data type
idInstance (I _{GS})	The entry identifier	Integer
name (N _{GS})	The entry name	String
lat (A _{GS})	The entry latitude	Double
lon (O _{GS})	The entry longitude	Double
idType (T _{GS})	GeoType code - Foreign Key (FK) for GeoType.idType	Integer
adminId1 (A1 _{GS})	First-order division - FK for GeoInstance.idInstance	Integer
adminId2 (A2 _{GS})	Second-order division - FK for GeoInstance.idInstance	Integer

Tabela 13 – Atributos da classe *GeoType* do esquema global.

Attribute name	Description	Data type
idType	The entry identifier	Integer
name	The entry name	String
description	The entry description	String
parentType	The entry parent (broader term) - FK for GeoType.idType	Integer

Tabela 14 – Fragmento do conjunto de instâncias de referência.

I _{GS}	N _{GS}	A _{GS}	O _{GS}	T _{GS}	A1 _{GS}	A2 _{GS}
175	Galapagos Islands	0.0	-90.5	4	73	-
52	Alps	46.4166667	10.0	15	165	-
149	Atlantic Ocean	10.0	-25.0	9	-	-
90	Niagara Falls	43.083416155	-79.06627052	21	123	-
16	Pão de Açúcar	-22.9472	-43.1561	14	101	-
34	Mississippi River	29.1510582	-89.2533842	19	109	-

Neste experimento, para detectar as co-ocorrências, foram utilizados procedimentos de comparação simples. Para atributos do tipo cadeias de caracteres (*string*) foi utilizado um operador de comparação de cadeias de caracteres. Para atributos numéricos, como latitude e longitude, primeiramente os valores foram truncados para então serem comparados, devido às diferenças nos sistemas de numeração utilizados. Futuros experimentos podem utilizar métodos de comparação mais sofisticados para obterem melhores resultados.

Como resultado, foram obtidas duas matrizes de ocorrências, apresentadas na Figura 23 e Figura 24. A Figura 23 mostra a matriz de ocorrências resultante da contagem de valores recorrentes entre as instâncias resultantes do ADL Gazetteer e as instâncias de referência. Já a Figura 24 mostra a matriz de ocorrências resultante da contagem de valores recorrentes entre as instâncias resultantes do GeoNames e as instâncias de referência. Note que, na matriz apresentada na Figura 24, mesmo com apenas 36 instâncias de

referência, existem 551 co-ocorrências dos valores do atributo N_{GS} (N_{GS} representa o atributo *name* do esquema global, vide Tabela 12) nos valores do atributo *name* nas instâncias coletadas do GeoNames. Isto se dá devido à operação de busca do Web Service localizar todas as instâncias onde ocorrem o nome enviado na consulta. Assim, para cada nome de instância de referência, teremos diversas ocorrências de instâncias nas fontes contendo o mesmo nome. Por exemplo, ao submeter para o GeoNames o nome da instância de referência “Mount Everest”, foram obtidas seis instâncias onde o valor “Mount Everest” co-ocorre (vide Tabela 15).

	I_{GS}	N_{GS}	A_{GS}	O_{GS}	T_{GS}	$A1_{GS}$	$A2_{GS}$
<i>identifier</i>	0	0	0	0	0	0	0
<i>placeStatus</i>	0	0	0	0	0	0	0
<i>name</i>	0	459	0	0	0	0	0
<i>displayName</i>	0	352	0	0	0	0	0
<i>footprintX</i>	0	0	14	134	0	0	0
<i>footprintY</i>	0	0	94	12	0	0	0
<i>class</i>	0	0	0	0	0	0	0
<i>thesaurus</i>	0	0	0	0	0	0	0
<i>names</i>	0	435	0	0	0	0	0
<i>relationships</i>	0	24	0	0	0	0	0

Figura 23 – Matrizes de ocorrências ADL Gazetteer X Esquema global.

	I_{GS}	N_{GS}	A_{GS}	O_{GS}	T_{GS}	$A1_{GS}$	$A2_{GS}$
<i>geonameId</i>	0	0	0	0	0	0	0
<i>lat</i>	0	0	188	15	0	0	0
<i>lon</i>	0	0	12	269	0	0	0
<i>name</i>	0	551	0	0	0	0	0
<i>countryCode</i>	0	0	0	0	0	0	0
<i>countryName</i>	0	0	0	0	0	0	0
<i>fcl</i>	0	0	0	0	0	0	0
<i>fcode</i>	0	0	0	0	0	0	0
<i>fclName</i>	0	0	0	0	0	0	0
<i>fcodeName</i>	0	0	0	0	0	0	0
<i>population</i>	0	0	0	0	0	0	0
<i>alternateName</i>	0	156	0	0	0	0	0
<i>elevation</i>	0	0	0	0	0	0	0
<i>adminCode1</i>	0	0	0	0	0	0	0
<i>adminCode2</i>	0	0	0	0	0	0	0
<i>adminName1</i>	0	8	0	0	0	0	0
<i>adminName2</i>	0	1	0	0	0	0	0

Figura 24 – Matrizes de ocorrências GeoNames X Esquema global.

Tabela 15 – Co-ocorrência de “Mount Everest” nos resultados do serviço de busca do GeoNames.

geonameId	lat	lng	name	countryCode	fcode
1283416	27.9833	86.9333	Mount Everest	NP	MT
1004850	-28.15	29.16667	Mount Everest	ZA	MT
4122419	33.78733	-93.3804	Mount Everest Church	US	CH
4334114	29.94326	-90.0904	Mount Everest Baptist Church	US	CH
4341122	29.94104	-90.089	Second Mount Everest Baptist Church	US	CH
4694788	32.70374	-96.7881	Greater Mount Everest Baptist Church	US	CH

	I_{GS}	N_{GS}	A_{GS}	O_{GS}	T_{GS}	$A1_{GS}$	$A2_{GS}$
<i>identifier</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>placeStatus</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>name</i>	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>displayName</i>	0.0000	0.7669	0.0000	0.0000	0.0000	0.0000	0.0000
<i>footprintX</i>	0.0000	0.0000	0.0305	0.2919	0.0000	0.0000	0.0000
<i>footprintY</i>	0.0000	0.0000	0.2048	0.0261	0.0000	0.0000	0.0000
<i>class</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>thesaurus</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>names</i>	0.0000	0.9477	0.0000	0.0000	0.0000	0.0000	0.0000
<i>relationships</i>	0.0000	0.0523	0.0000	0.0000	0.0000	0.0000	0.0000

Figura 25 – Matriz de mapeamentos ADL Gazetteer X Esquema global.

	I_{GS}	N_{GS}	A_{GS}	O_{GS}	T_{GS}	$A1_{GS}$	$A2_{GS}$
<i>geonameId</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>lat</i>	0.0000	0.0000	0.2674	0.0213	0.0000	0.0000	0.0000
<i>lon</i>	0.0000	0.0000	0.0171	0.3826	0.0000	0.0000	0.0000
<i>name</i>	0.0000	0.7838	0.0000	0.0000	0.0000	0.0000	0.0000
<i>countryCode</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>countryName</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>fcl</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>fcode</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>fclName</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>fcodeName</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>population</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>alternateName</i>	0.0000	0.2219	0.0000	0.0000	0.0000	0.0000	0.0000
<i>elevation</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>adminCode1</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>adminCode2</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>adminName1</i>	0.0000	0.0114	0.0000	0.0000	0.0000	0.0000	0.0000
<i>adminName2</i>	0.0000	0.0014	0.0000	0.0000	0.0000	0.0000	0.0000

Figura 26 – Matriz de mapeamentos GeoNames X Esquema global.

Para descobrir os mapeamentos, é necessário aplicar o processo de normalização descrito na seção 4.1. Após serem aplicadas as fórmulas (3) e (4),

foram obtidas as matrizes de mapeamentos entre o esquema do ADL Gazetteer e o esquema global (Figura 25) e entre o esquema do GeoNames e o esquema global (Figura 26).

A partir destas matrizes, os mapeamentos são identificados pelos elementos cujo valor é maior do que 0.2. A Tabela 16 mostra os alinhamentos encontrados entre o esquema do ADL Gazetteer e o esquema global. A Tabela 17 mostra os alinhamentos encontrados entre o esquema do GeoNames e o esquema global.

Tabela 16 – Mapeamentos corretos entre ADL Gazetteer X Esquema global.

Global Schema	ADL Gazetteer
name (N _{GS})	name
name (N _{GS})	displayName
name (N _{GS})	names
lat (A _{GS})	footprintY
lon (O _{GS})	footprintX

Tabela 17 – Mapeamentos corretos entre GeoNames X Esquema global.

Global Schema	GeoNames
name (N _{GS})	name
name (N _{GS})	alternateName
lat (A _{GS})	lat
lon (O _{GS})	lon

Como esperado, os atributos de “tipo” não foram alinhados pois o esquema global foi modelado utilizando uma chave estrangeira da tabela GeoType para a tabela GeoInstance. Neste caso, o procedimento ideal seria criar uma visão do esquema global, contendo os nomes dos tipos, e então realizar os alinhamentos utilizando os elementos do esquema desta visão.

Outro exemplo de melhoria a ser feita neste experimento seria a modelagem do esquema global, ou a criação de uma visão dele, descartando os atributos identificadores, tais como: idInstance, adminIds, etc. Como se tratam de identificadores para controle local do banco de dados, eles jamais serão alinhados com nenhum identificador de um banco de dados externo, podendo assim serem descartados.

Dados estes problemas identificados durante a validação, o ideal é que seja criada uma visão sobre os esquemas a serem alinhados, utilizando apenas os atributos relevantes entre as fontes envolvidas.

De posse destes mapeamentos, um mediador, que utilize a abordagem de alinhamento *a priori*, pode receber consultas formuladas utilizando o esquema global e traduzi-las para os esquemas das fontes a serem consultadas. Por exemplo, se o mediador recebe a consulta Q utilizando atributos do esquema global S_G , ele pode facilmente traduzi-la para Q_1 e Q_2 , utilizando termos de S_A e S_B , respectivamente.

```
Q =  SELECT * FROM ALL
      WHERE geoinstance.name LIKE "%Mount Everest%"

QA = SELECT * FROM source1
      WHERE source1.name LIKE "%Mount Everest%"
      OR source1.displayName LIKE "%Mount Everest%"
      OR source1.names LIKE "%Mount Everest%"

QB = SELECT * FROM source2
      WHERE source2.name LIKE "%Mount Everest%"
      OR source2.alternateName LIKE "%Mount Everest%"
```

4.4.3

Validação e teste da abordagem adaptativa

Como detalhado na seção 4.3, na abordagem adaptativa para alinhamento de esquemas, diferentemente da abordagem *a priori* descrita na seção 4.2, o conjunto de instâncias de referência adotado é a coleção de instâncias retornada de uma fonte como resultado de uma consulta de usuário. Os valores dos atributos destas instâncias são verificados contra a coleção de instâncias retornadas da consulta à segunda fonte. Supõe-se consultas realizadas numa mesma seção de usuário à duas fontes distintas. Nesta abordagem são criados mapeamentos entre os esquemas das fontes a serem integradas.

Para conduzir os experimentos, foram utilizados os esquemas das fontes apresentadas na seção 4.4.1: GeoNames e ADL Gazetteer. Ambos os serviços foram consultados utilizando as operações de busca por nome de lugar.

Os elementos dos esquemas utilizados neste experimento são detalhados a seguir. A Tabela 18 mostra os elementos do esquema de exportação da fonte ADL Gazetteer utilizado neste experimento. Já a Tabela 19 mostra os elementos do esquema de exportação da fonte GeoNames utilizado neste experimento.

Tabela 18 – Esquema de exportação da fonte ADL Gazetteer.

Attribute name	Description	Data type
identifier (c ₁)	Entry local id	String
gnis_identifier (c ₂)	Entry id on GNIS	String
placeStatus (c ₃)	Entry place-status (current or former)	String
displayName (c ₄)	Display name	String
names (c ₅)	Alternative names	names
bounding-box_X (c ₆)	Entry longitude	Number
bounding-box_Y (c ₇)	Entry latitude	Number
ftt_class (c ₈)	Entry class of FTT	String
gnis_class(c ₉)	Entry class of GNIS	String

Tabela 19 – Esquema de exportação da fonte GeoNames.

Attribute name	Description	Data type
name (d ₁)	Primary name	String
lat (d ₂)	Latitude	Number
lng (d ₃)	Longitude	Number
geonameId (d ₄)	Identifier	String
countryCode (d ₅)	Country code (ISO-3166 2-letter code)	String
countryName (d ₆)	Country name	String
fcl(d ₇)	Feature type super class code	String
fcode (d ₈)	Feature type classification code	String
fclName (d ₉)	Feature type super class name	String
fcodeName (d ₁₀)	Feature type classification name	String
population (d ₁₁)	Population	Number
alternateNames (d ₁₂)	Alternative names	String
elevation (d ₁₃)	Elevation, in meters	Number
adminCode1 (d ₁₄)	Code for 1st adm. division	String
adminName1 (d ₁₅)	Name for 1st adm. division	String
adminCode2 (d ₁₆)	Code for 2nd adm. division	String
adminName2 (d ₁₇)	Name for 2nd adm. division	String
timezone (d ₁₈)	Timezone description	String

Para este experimento, foi submetida a consulta pela palavra chave “Alps”. A primeira fonte consultada foi a ADL Gazetteer, a qual retornou 71 entradas. Já a fonte GeoNames, retornou 77 entradas. De posse de ambos os conjuntos de resultados em cache, é gerada a matriz de ocorrências. A matriz de ocorrências contém, para cada par de atributos, a contagem de co-ocorrência dos valores de objetos do conjunto de resultado da primeira consulta nos valores de objetos do conjunto de resultado da segunda consulta. A Figura 27 mostra a matriz de ocorrências gerada. De posse da matriz de ocorrências, foi aplicada a fórmula

(5) apresentada na seção 4.3, gerando a matriz de mapeamentos, apresentada na Figura 28.

A partir da matriz de mapeamentos, os mapeamentos corretos são identificados pelos elementos cujo valor for maior que o limiar de 0.2. A

Tabela 20 mostra os mapeamentos corretos encontrados entre o esquema do ADL Gazetteer e o GeoNames.

	c ₁	c ₂	c ₃	c ₄	c ₅	c ₆	c ₇	c ₈	c ₉
d ₁	0	0	0	0	77	0	0	0	0
d ₂	0	0	0	0	0	0	61	0	0
d ₃	0	0	0	0	0	67	0	0	0
d ₄	0	0	0	0	0	0	0	0	0
d ₅	0	0	0	0	0	0	0	0	0
d ₆	0	0	0	0	0	0	0	0	0
d ₇	0	0	0	0	0	0	0	0	0
d ₈	0	0	0	0	0	0	0	0	55
d ₉	0	0	0	0	0	0	0	3	0
d ₁₀	0	0	0	0	0	0	0	76	0
d ₁₁	0	0	0	0	0	0	0	0	0
d ₁₂	0	0	0	0	33	0	0	0	0
d ₁₃	0	0	0	0	0	0	0	0	0
d ₁₄	0	0	0	0	0	0	0	0	0
d ₁₅	0	0	0	0	0	0	0	0	0
d ₁₆	0	0	0	0	0	0	0	0	0
d ₁₇	0	0	0	0	0	0	0	0	0
d ₁₈	0	0	0	0	0	0	0	0	0

Figura 27 – Matriz de ocorrências GeoNames X ADL Gazetteer.

Tabela 20 – Mapeamentos corretos entre GeoNames X ADL Gazetteer.

ADL Gazetteer	GeoNames
names (c ₅)	name (d ₁)
names (c ₅)	alternateNames (d ₁₂)
bounding-box_X (c ₆)	lng (d ₃)
bounding-box_Y (c ₇)	lat (d ₂)
ftt_class (c ₈)	fcodeName (d ₁₀)
gnis_class(c ₉)	fcode (d ₈)

	c ₁	c ₂	c ₃	c ₄	c ₅	c ₆	c ₇	c ₈	c ₉
d ₁	0	0	0	0	0.9873	0	0	0	0
d ₂	0	0	0	0	0	0	0.7822	0	0
d ₃	0	0	0	0	0	0.8591	0	0	0
d ₄	0	0	0	0	0	0	0	0	0
d ₅	0	0	0	0	0	0	0	0	0
d ₆	0	0	0	0	0	0	0	0	0
d ₇	0	0	0	0	0	0	0	0	0
d ₈	0	0	0	0	0	0	0	0	0.7053
d ₉	0	0	0	0	0	0	0	0.0386	0
d ₁₀	0	0	0	0	0	0	0	0.9746	0
d ₁₁	0	0	0	0	0	0	0	0	0
d ₁₂	0	0	0	0	0.4232	0	0	0	0
d ₁₃	0	0	0	0	0	0	0	0	0
d ₁₄	0	0	0	0	0	0	0	0	0
d ₁₅	0	0	0	0	0	0	0	0	0
d ₁₆	0	0	0	0	0	0	0	0	0
d ₁₇	0	0	0	0	0	0	0	0	0
d ₁₈	0	0	0	0	0	0	0	0	0

Figura 28 – Matriz de mapeamentos GeoNames X ADL Gazetteer.

4.5

Considerações

Este capítulo apresentou as abordagens para alinhamento de esquemas conceituais simples utilizando instâncias.

A seção 4.2 introduziu a abordagem *a priori* para alinhamento de esquemas. Esta abordagem baseia-se na definição de um esquema global e na coleta de um conjunto de instâncias de referência para servirem como insumos para consultas submetidas às fontes a serem integradas. As instâncias retornadas nos resultados dessas consultas servem como evidências para os alinhamentos entre os esquemas das fontes e o esquema global. Porém, como o processo de definição do esquema global e de coleta das instâncias de referência pode ser trabalhoso, foi proposta uma abordagem adaptativa, descrita na seção 4.3. A abordagem adaptativa para alinhamento de esquemas conceituais simples usa como evidências as instâncias retornadas a partir de consultas realizadas por usuários.

Nas seções 4.4.2 e 4.4.3 foram exemplificados casos de heterogeneidade entre esquemas conceituais de bancos de dados geográficos reais: do Geonames e do ADL Gazetteer, para validação e teste das abordagens *a priori* e adaptativa, as quais apresentaram bons resultados.