Aplikasi Pengolah Bahasa Alami untuk Query Basisdata XML

Sri Hartati
Elektronika dan Instrumentasi, FMIPA, Universitas Gadjah Mada

Eri Zuliarso
Teknik Informatika, FTI, Universitas Stikubank Semarang
email : erizuliarso@yahoo.com

Abstrak : Telah dikembangkan sebuah aplikasi pengolah bahasa alami untuk melakukan query basisdata XML. Data yang tersimpan dalam basisdata XML berupa bibliografi koleksi perpustakaan. Bahasa alami yang diproses untuk query merupakan bahasa Indonesia sehari-hari yang mengikuti pola kalimat tertentu sesuai dengan tata bahasa Indonesia.

Query basisdata XML dilakukan setelah aplikasi ini menerima masukan berupa kalimat bahasa Indonesia yang sederhana. Kalimat ini mengikuti pola aturan produksi yang telah ditetapkan dan mengikuti tatabahasa Indonesia. Setelah proses penginputan kalimat bahasa alami selesai, dilakukan proses analisis leksikal, sintaks, parsing, dan semantiks untuk verifikasi bahasa alami yang masuk. Hasil yang benar akan diterjemahkan menjadi XQuery sebagai suatu perintah untuk melakukan query pada basisdata XML. Perintah ini kemudian akan diproses untuk menghasilkan keluaran hasil operasi query tersebut.

Hasil penelitian menunjukkan bahwa aplikasi pengolahan bahasa alami melakukan query pada basisdata XML dengan benar, dan memberi kemudahan bagi user untuk melakukan operasi pencarian informasi bibliografi.

Kata kunci : bahasa alami, natural language processing, basisdata XML, XQuery

PENDAHULUAN

XML muncul sebagai standar yang dominant untuk representasi dan pertukaran data melalui Internet. XML mempunyai struktur bersarang dan memaparkan dirinya, hal ini memudahkan bagi aplikasi untuk memodelkan dan bertukar data. Dengan banyaknya data yang direpresentasikan sebagai dokumen XML, menjadi sangat perlu untuk menyimpan dan meng-query dokumen-dokumen XML ini. Untuk mengatasi masalah ini, telah dibangun system database XML.

Dengan adanya system basisdata XML ini memang memberikan kemudahan untuk menyimpan dan meng-query data yang mempunyai format dokumen atau semi terstruktur. Namun demikian dari sisi pemakai, dipaksa untuk belajar bagaimana menggunakan system tersebut. Baik itu pada saat membangun tabel data maupun pada saat akan melakukan query. Hal ini dikarenakan format XML mempunyai bahasa query sendiri yaitu XQuery. XQuery mempunyai struktur tata bahasa yang cukup rumit, hal ini mungkin akan memberikan kesulitan bagi para pemakai biasa.

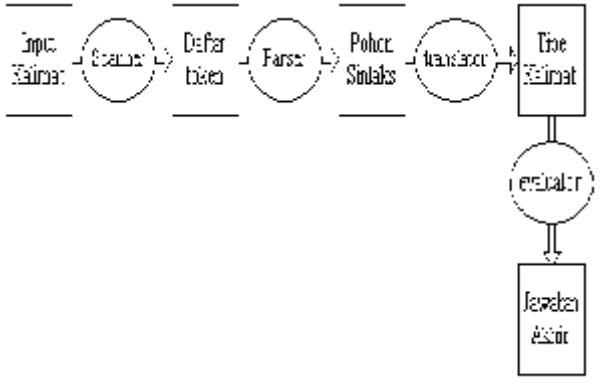
Untuk itu antarmuka dengan bahasa alami adalah suatu jawaban yang benar, khususnya untuk pengguna awam dan pengguna akhir. Antarmuka bahasa alami ke basisdata adalah sebuah system yang membolehkan pemakai untuk mengakses informasi yang tersimpan dalam basisdata dengan mengetikkan permintaan yang diekspresikan dalam bahasa alami.

Penelitian di bidang Antarmuka bahasa alami ke basisdata sudah banyak dilakukan. Shan Wang,2000 mengembangkan system query berbahasa alami Cina untuk mengakses database yang diberi nama Nchiql. Nchiql didisain sebagai system query ke system basisdata Cina dan dapat digunakan ke aplikasi yang berbeda pada DBMS yang berbeda. Andayani, 2002, telah melakukan penelitian Query Bahasa Indonesia untuk basis data akademik. Database yang digunakan adalah relasional. Aplikasi pemrosesan bahasa alami untuk query basis data ini menggunakan input dalam bahasa Indonesia. Yunyao Li,2005 mengembangkan sebuah

antarmuka bahasa alami untuk meng-query database XML. NaLIX adalah sebuah antarmuka query bahasa alami interaktif yang berlaku umum ke basisdata XML.

Dalam tulisan ini akan dibahas aplikasi pemrosesan bahasa alami sebagai pengganti query pada basis data XML dengan menggunakan bahasa Indonesia. Aplikasi pemrosesan bahasa alami yang dibuat menggunakan bahasa Indonesia sebagai interpreter query pada basis data bibliografi yang menyimpan informasi tentang bibliografi koleksi perpustakaan.

STRUKTUR PENGOLAH BAHASA ALAMI



Gambar 1: Komponen pengolah bahasa alami untuk Aplikasi Pengolah Bahasa Alami Untuk Query Basisdata XML

Komponen pengolah bahasa alami untuk Aplikasi Pengolah Bahasa Alami Untuk Query Basisdata XML ditunjukkan pada Gb.1. Setiap kalimat bahasa alami, berupa kalimat berbahasa Indonesia, yang dimasukkan akan melewati proses yang dilakukan oleh scanner, parser, translator dan evaluator sebelum mendapatkan jawaban akhir. Scanner akan melakukan pemeriksaan bentuk kalimat dan mengelompokkannya menjadi daftar token yang kemudian diteruskan ke proses berikutnya yang dilakukan oleh parser. Dalam proses ini parser melakukan pelacakan terhadap token-token tersebut untuk dibandingkan dengan daftar token yang telah ditetapkan. Tanslator akan menterjemahkan hasil parsing untuk mengecek kesesuaian struktur kalimat dengan pola atau aturan produksi yang. Selanjutnya, hasil proses yang sesuai dengan pola kalimat ini akan diteruskan ke evaluator.

XML, XML SCHEMA DAN BASIS DATA XML

Sebuah bahasa markup adalah sebuah mekanisme untuk mengidentifikasi struktur dalam sebuah dokumen. Spesifikasi XML mendefinisikan suatu cara standard untuk menambahkan markup ke dokumen-dokumen[Klein, 2001].

Sintak XML adalah bagian dari standar pengolehan teks internasional SGML yang secara spesifik dimaksudkan untuk digunakan di web. XML tidak ada yang memiliki, dapat di validasi, dapat dibaca manusia dan mempunyai kemampuan untuk merepresentasikan struktur yang komplek[Bray,2004]. Hal terpenting dari XML adalah bahwa pustaka dari penanda dan kombinasiny tidak tetap, tetapi dapat didefinisikan untuk tiap aplikasi.

XML Schema adalah bahasa yang digunakan untuk mendefinisikan struktur dokumen XML. Salah satu karakterisitik dari XML Schema adalah sintaknya didasarkan pada XML itu sendiri. Hal ini memudahkan untuk dibaca, dan tidak diperlukan parser, editor dan lain-lain yang berbeda dengan yang digunakan untuk XML.

Bourret[2005] Basisdata XML adalah sistem perangkat lunak yang digunakan untuk menyimpan data yang membolehkan data untuk diimpor, diakses dan diekspor dalam format XML. Basisdata XML mempunyai keunggulan lebih baik dibandingkan dengan system basisdata relasional jika data yang akan disimpan berupa dokumen. Dengan basisdata XML juga memungkinkan untuk melakukan penelusuran isi dokumen. O'Connell (2005, 9.2) memberikan suatu alasan penggunaan XML dalam basisdata, yaitu dengan meningkatnya penggunaan XML secara umum untuk transport data, yang berarti data diekstrak dari basisdata dan diubah menjadi dokumen XML dan sebaliknya. Akan lebih efisien (dalam pemikiran biaya konversi) dan lebih mudah untuk menyimpan data dalam format XML

XML QUERY LANGUAGE (XQUERY)

XQuery adalah sebuah bahasa untuk melakukan query terhadap data XML. XQuery didisain dari XML Query Working Group dari W3C dengan tujuan tunggal untuk meng-query data yang tersimpan dalam format XML[*Walmsley,2007]*. Hal yang penting tentang XQuery adalah bahwa XQuery bekerja dengan semua dokumen XML, apakah dokumen tersebut bertipe, tidak bertipe atau kombinasi dari keduanya.

XQuery bekerja dengan semua dokumen XML, apakah dokumen tersebut bertipe, tidak bertipe atau kombinasi dari keduanya. Hal ini dapat dilakukan dengan menggunakan fungsi navigasi XPath.

ARSITEKTUR SYSTEM

Sistem terdiri dari 3 lapisan yaitu :

a. lapisan database

   Lapisan ini digunakan untuk menyimpan dokumen XML. Dalam system ini digunakan DBMS SQL Server 2005. SQL Server 2005 mengenalkan tipe data XML. Tipe data ini dapat digunakan dalam definisi tabel untuk mendefinisikan tipe sebuah kolom, tipe variabel dalam kode prosedural Transact-SQL, dan sebagai parameter prosedur. Kolom, variabel dan parameter dari tipe data XML dapat dibatasi dengan XML Schema. XML Schema didefinisikan dalam katalog SQL Server.

b. lapisan bahasa query

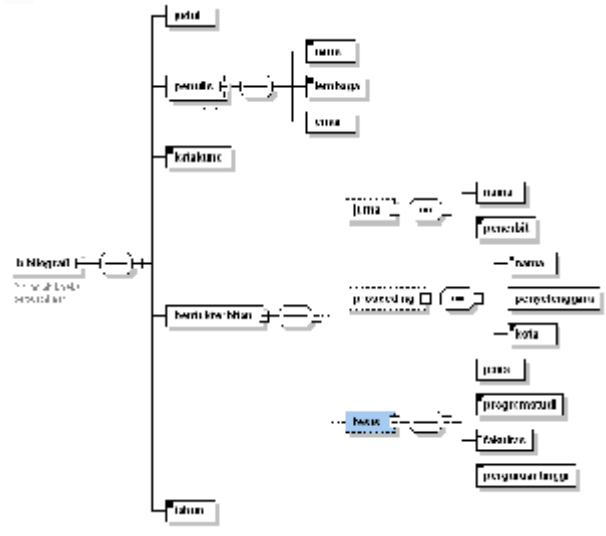
   XML, seperti basisdata relasional, mempunyai bahasa query sendiri yang dioptimasi untuk format data. Untuk SQL Server 2005, Microsoft telah menambahkan dukungan *server-side untuk XQuery. Berbasis pada bahasa query XPath, XQuery adalah bahasa yang dapat meng-query data XML terstruktur dan semi-terstruktur.* Berpasangan dengan tipe data xml, hal ini mempercepat dan mengefisienkan penyimpanan dan temukembali data XML.

c. lapisan aplikasi

   Lapisan ini merupakan antarmuka menggunakan bahasa Indonesia. Bahasa pemrograman yang digunakan adalah Java. Java menyediakan banyak fasilitas yang memudahkan untuk mengimplementasikan system yang dibuat

STRUKTUR DATA

Data yang disimpan dalam basisdata XML berupa bibliografi koleksi perpustakaan. Bibliografi (bibliography) adalah daftar buku-buku dan karya ilmiah lain seperti artikel jurnal yang tersusun secara sistematis. Adapun struktur masukan untuk bibliografi digambarkan dalam skema pada gambar

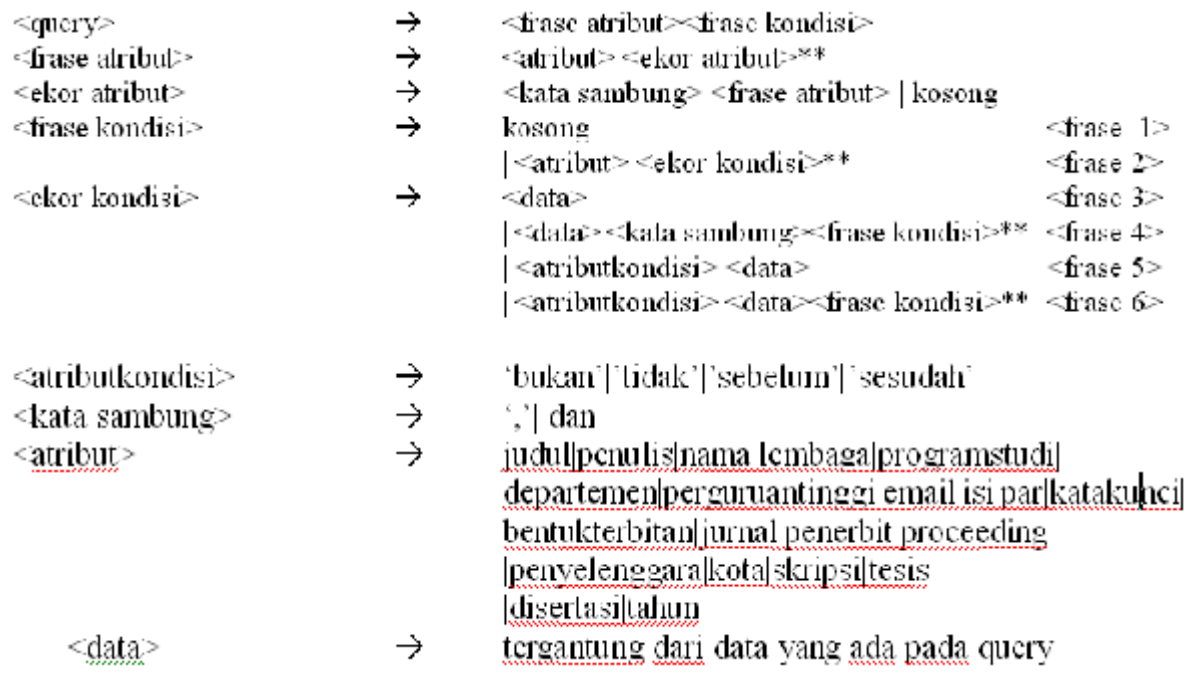


Gambar 2. XML Schema bibliografi

PENYUSUNAN ATURAN PRODUKSI

Aplikasi pemrosesan bahasa alami untuk query basis data ini menggunakan input dalam bahasa Indonesia. Input berkisar pada pertanyaan-pertanyaan untuk mengakses informasi atau data dari basisdata. Dengan demikian, meskipun bahasa Indonesia sudah mempunyai grammar dan aturan produksi, akan tetapi dalam aplikasi ini perlu ditentukan aturan produksi yang secara khusus menangani pola-pola pertanyaan pada input.

Berdasarkan pola keteraturan pertanyaan-pertanyaan, maka aturan produksi dengan symbol awal <query> ditentukan seperti dalam gambar :



Gambar 3. Struktur Grammar

Berdasarkan contoh-contoh pertanyaan input dalam aplikasi ini, ada 7 tipe query yang diidentifikasi. Setiap tipe mempunyai notasi dan pola input yang berbeda. Notasi tipe query akan digunakan lebih lanjut dalam proses implementasi. Pola input menentukan penggolongan tipe query berdasarkan input yang sesuai dengan pola tersebut.

Simbol-simbol yang digunakan dalam penulisan pola input adalah sebagai berikut :

[T/P] : adalah kata Tanya atau kata perintah yang mengawali input, yang dalam proses selanjutnya dapat diabaikan. Kata tersebut diapit tanda [] yang berarti bersifat opsi.

[Plk] : satu atau lebih kata-kata pelengkap.

<atribut> : nama atribut yang terdapat dalam daftar token, atau kata-kata yang merupakan sinonim atribut.

<bukan> : kata-kata yang mempunyai arti sama dengan "bukan"

<opr> : symbol atau kata-kata yang berfungsi sebagai operator

<int> : kata-kata yang menjelaskan tentang intensitas, yang dibedakan menjadi maks dan min.

Ketujuh tipe query yang diidentifikasi adalah sbb:

a. Tipe q_a (query-atribut)

Tipe query yang hanya berisi satu atribut yang akan ditampilkan. Query ini merupakan tipe yang paling sederhana, yang hanya memuat atribut yang ditanyakan.

Tipe ini mempunyai pola input, yaitu :

i. [T/P] <atribut> [Plk]

Contoh : Apa judul bibliografi

ii. [T/P] [Plk] <atribut>

Contoh : Siapa yang menjadi penulis

b. Tipe q_aa (query-atribut-atribut)

Tipe query yang berisi beberapa atribut yang akan ditampilkan. Tipe query ini memuat beberapa atribut yang ditanyakan. Untuk memisahkan satu atribut dengan atribut berikutnya digunakan kata sambung 'dan' atau tanda baca koma ','.

Tipe ini mempunyai pola input, yaitu :

i. [T/P] <atribut> (<ktsambung> <atribut>)*

Contoh : Apa judul , penulis dan lembaga

ii. ([T/P] [Plk]* <atribut>)*

Contoh : Apa judul dan siapa yang menjadi penulis

c. Tipe q_a_opr (query-atribut-atribut-operator)

Tipe query berisi satu atribut yang akan ditampilkan dan satu kondisi.

Tipe ini mempunyai pola input, yaitu :

i. [T/P] <atribut> [Plk] <atribut> <data>

Contoh : cari judul dengan kata kunci natural language

ii. [T/P] <atribut> [Plk] <atribut> <data>

Contoh : tampilkan judul dengan judul mengandung alami

d. Tipe q_a_opr (query- atribut-atribut-atribut-operator- atribut-operator)

Tipe query berisi beberapa atribut yang akan ditampilkan dan beberapa kondisi.

Tipe query ini memuat beberapa atribut yang ditanyakan. Untuk memisahkan satu atribut dengan atribut berikutnya digunakan kata sambung 'dan' atau tanda baca komma ','. Demikian juga untuk memisahkan satu atribut kondisi dengan atribut kondisi berikutnya digunakan kata sambung 'dan' atau tanda baca komma ','.

Tipe ini mempunyai pola input, yaitu :

[T/P] <atribut> (<ktsambung> <atribut>)*
[Plk] <atribut> <data> <ktsambung> <atribut><data>

Contoh :

i. Tampilkan judul dan penulis dengan penulis Sri Hartati dan tahun 2005

ii. Tampilkan judul dan penulis di jurnal Tekno dan penulis Sri dan judul Citra

e. Tipe q_operator (query-atribut-operator-<data>)

Tipe query berisi beberapa atribut yang akan ditampilkan dan kondisi operator 'sebelum' atau 'sesudah'.

Tipe ini mempunyai pola input, yaitu :

[T/P] <atribut> (<ktsambung> <atribut>)*
[Plk] <atribut><op_tahun><data>

Contoh :

i. Judul dan penulis dengan tahun sebelum 2006

ii. Judul dan penulis dengan tahun sesudah 2005

f. Tipe q_bukan (query-atribut-bukan-data)

Tipe query berisi beberapa atribut yang akan ditampilkan dan kondisi operator "bukan" atau "tidak" atau "selain".

Tipe ini mempunyai pola input, yaitu :

[T/P] <atribut> (<ktsambung> <atribut>)*
[Plk] <atribut><op_bukan><data>

Contoh :

i. Judul dan penulis dengan penulis bukan Sri

g. Tipe q_a_bukan (query-atribut-atribut-bukan-data)

Tipe query berisi beberapa atribut yang akan ditampilkan dan kondisi operator "bukan" atau "tidak" atau "selain".

Tipe ini mempunyai pola input, yaitu :

[T/P] <atribut> (<ktsambung> <atribut>)*
[Plk] <atribut><op_bukan><data>

Contoh :

i. Judul dan penulis dengan penulis bukan Sri

HASIL UJI COBA

Dari aturan produksi yang telah ditetapkan maka sistem hanya bisa menerima masukan yang sesuai, misalkan bentuk kalimat perintah "Siapa yang menjadi penulis". Proses awal yang dilakukan terhadap kalimat tersebut adalah pembentukkan daftar token yang dilakukan oleh scanner. Token-token ini akan diproses oleh parser. Parser melakukan pelacakan terhadap pembentukan kalimat yang kemudian dianalisa kesesuaiannya dengan

aturan produksi yang ada. Penterjemahan kalimat hasil dari pohon sintaks dilakukan oleh translator yang menghasilkan tipe kalimat. Dalam proses ini akan diketahui apakah kalimat perintah yang dimasukkan itu sesuai dengan aturan produksi yang ditetapkan atau tidak untuk mendapatkan jawaban akhir yang diinginkan user. Bila sesuai, maka tipe kalimat diproses oleh evaluator untuk mendapatkan hasil operasi query.

Beberapa contoh hasil pengujian disajikan untuk melihat beberapa hasil operasi query pada basis data XML. Dalam hasil uji coba ditampilkan masukan dalam bahasa Indonesia, statemen XQuery dan hasil evaluasi XQuery pada basisdata XML pada SQL Server.

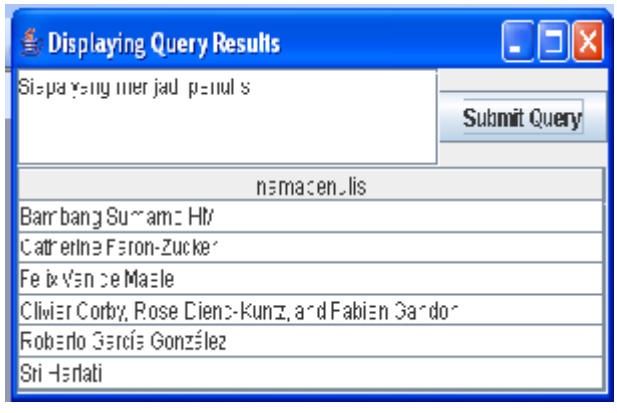
a. Pengujian 1

Query input :

Siapa yang menjadi penulis

Statemen XQuery :

Select distinct a.value('nama','varchar(200)') as namapenulis from perpustakaan CROSS APPLY isibibliografi.nodes('/bibliografi/penulis') as result(a)



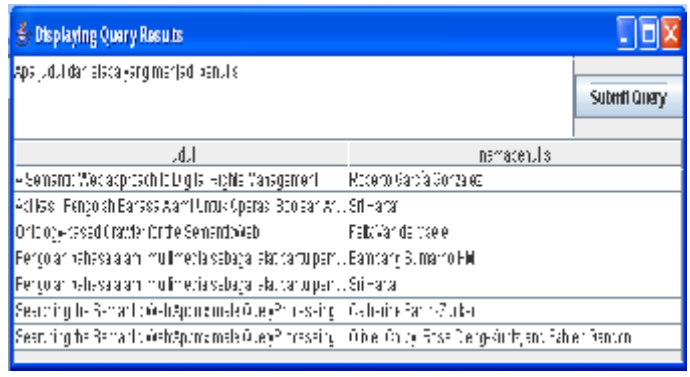
Gambar 4. Hasil dari Query Pengujian 1

b. Pengujian 2

Query input : Apa judul dan siapa yang menjadi penulis

Statemen XQuery :

Select distinct isibibliografi.value('/bibliografi[1]/judul[1]', 'varchar(200)') as judul, a.value('nama','varchar(200)') as namapenulis from perpustakaan CROSS APPLY isibibliografi.nodes('/bibliografi/penulis') as result(a)



Gambar 5. Hasil Query Pengujian 2

KESIMPULAN

Hasil penelitian menunjukkan bahwa sistem mampu menterjemahkan kalimat bahasa Indonesia ke bahasa XQuery, sehingga aplikasi ini dapat melakukan query pada basisdata XML.

Sistem yang dibuat telah berhasil memproses kalimat berdasarkan jenis query yang umum pada suatu bibliographi.

DAFTAR PUSTAKA

1. [And2002] Andayani, Sri, 2002, Qubin : Query Bahasa Indonesia untuk basis data akademik, Tesis magister Komputer, Program Pasca sarjana Ilmu Komputer, UGM
2. [Bel2003] Belanger, Terry. "Descriptive Bibliography" Bibliographical Society of America, 2003. Excerpted from Jean Peters, ed., Book Collecting: A Modern Guide (New York and London: R. R. Bowker, 1977), 97-101.
3. [Bea2004] Beauchemin, Bob, 2004 ,*AFirst Look at SQL Server 2005 for Developers,* Addison-Wesley,
4. [Bou] Bourret, Ronald ,http://www.rpbourret.com
5. [Hor2005] Horton, Ivor ,Beginning Java™ 2, JDK™ 5 Edition, Wiley Publishing, Inc., Indianapolis, Indiana, 2005

6. [Klei2006] Klein, Scott ,Professional SQL Server™ 2005 XML, Wiley Publishing, Inc., Indianapolis, Indiana, 2006
7. [Wan200] Wang, Shan,Nchiql: A Chinese Natural Language Query System to Databases, Proceedings of the 1999 International Symposium on Database Applications in Non-Traditional Environments,IEEE, 2000
8. [Tso1995] Tsopoulos, I, Natural Language Interface, Journal of Natural languages Engineering, Cambridge University Press., 1995
9. [Yun2005] Yunyao Li, NaLIX: an Interactive Natural Language Interface for Querying XML, *SIGMOD*, Baltimore, Maryland, USA, 2005
10. [W3C2000] World Wide Web Consortium, "Extensible Markup Language (XML) 1.0 (Second Edition)", W3C Recommendation, October 2000.
11. [W3C2000] World Wide Web Consortium, "XQuery: A Query Language for XML", W3C Working Draft, February 2000.