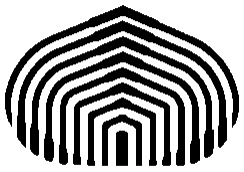


UNIVERSIDAD SIMÓN BOLÍVAR
DECANATO DE ESTUDIOS PROFESIONALES
COORDINACIÓN DE INGENIERÍA DE LA COMPUTACIÓN

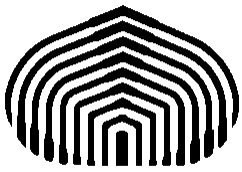
EXTRACCIÓN DE INFORMACIÓN FOCALIZADA BASADA EN RESPUESTAS INCOMPLETAS (FIE)

Por:
FRANCISCO RODRÍGUEZ DRUMOND

Realizado con la asesoría de:
PROF. SORAYA ABAD

PROYECTO DE GRADO
Presentado ante la Ilustre Universidad Simón Bolívar
como requisito parcial para optar al título de
Ingeniero de Computación

Sartenejas, Septiembre de 2012



UNIVERSIDAD SIMÓN BOLÍVAR
DECANATO DE ESTUDIOS PROFESIONALES
COORDINACIÓN DE INGENIERÍA DE LA COMPUTACIÓN

ACTA FINAL PROYECTO DE GRADO

EXTRACCIÓN DE INFORMACIÓN FOCALIZADA BASADA EN RESPUESTAS INCOMPLETAS (FIE)

Presentado por:
FRANCISCO RODRÍGUEZ DRUMOND

Este Proyecto de Grado ha sido aprobado por el siguiente jurado examinador:

_____
PROF. 1

_____
PROF. 2

_____
PROF. 3

Sartenejas, X/X/12

Resumen

RESUMEN GOES HERE - To be done.

Acrónimos

FIE	Focalized Information Extraction	Extracción Focalizada de Información
DIA	Document Interrogation Architecture	Arquitectura de Interrogación de Documentos

Índice general

Índice de tablas

Índice de figuras

Introducción

Este trabajo consiste en el estudio de una variante del problema de enrutamiento de vehículos (VRP por las siglas de Vehicle Routing Problem) el cual se define como un problema de optimización combinatoria del área de logística y distribución de bienes.

Capítulo 1

Planteamiento del Problema

Plantamiento del problema goes here.

Capítulo 2

Marco Teórico

Marco Teórico goes here.

2.1. Revisión de trabajos existentes en el área de extracción de Información

En "A survey of Uncertain Data Algorithms and Applications" (2005) [1], Aggarwal y Fellow presentan un estudio de las tecnologías que se han desarrollado para trabajar con datos inciertos. Los autores definen los datos inciertos como datos que pueden tener estar incompletos o que pueden contener errores. Por ejemplo, se puede tener una base de datos con datos de mediciones meteorológicas cuya precisión depende de los instrumentos utilizados y que por ende sus datos pueden contener errores. Puede darse también el caso en el que no se tenga la información completa, como en la base de datos un censo en la cual no se tiene la información de todos los ciudadanos de un país. En el caso particular de este proyecto de grado se trabaja con datos incompletos.

En base a la definición de datos inciertos, los autores examinan las técnicas más recientes en tres campos fundamentales: modelado de datos inciertos, manejo de datos inciertos y minería de datos inciertos. Este proyecto de grado está enmarcado en las dos primeras categorías: modelado y manejo de datos inciertos. Esto se debe a que se quiere estudiar la forma como modelar el problema así como un conjunto de mecanismos que permitan completar la información que falta mediante extracción focalizada.

Los autores definen una base de datos probabilística como un espacio de probabilidad finito en el cual los resultados son las posibles bases de datos consistentes con un esquema dado. En pocas palabras, se tiene una representación de los "posibles mundos". Existen varias soluciones para ello, como modelar la probabilidad de que una tupla esté en la base de datos o, si se quiere más detalle, la probabilidad de que un atributo de una tupla de base de datos esté presente. Sobre esta base se pueden construir técnicas para

el manejo de datos y para realizar minería sobre ellos.

Por otro lado, Aggarwal y Fellow (2006) [1] proponen utilizar una función de probabilidad sobre la correctitud de los datos presentes en la base de datos. Si bien el objetivo de este proyecto es el desarrollo de un mecanismo para contestar una consulta cuya respuesta no está en su totalidad en la ontología, la aproximación que hacen estos autores puede ser útil para definir una medida de la calidad de los datos presentes en la Base de Datos.

Es importante destacar sin embargo, una referencia que se hace al trabajo "Evaluating Probabilistic Queries over Imprecise Data" (2003), de Cheng, Kalashnikov y Prabhakar [2]. En el mismo se hace un análisis sobre las consultas con datos imprecisos o propensos a errores. Una vez más, esto es diferente de lo que se quiere en este trabajo de investigación: trabajar consultas incompletas. Sin embargo, Cheng et al proponen una clasificación de consultas probabilísticas que puede ser tomada en cuenta para clasificar las consultas incompletas.

Básicamente los autores toman en consideración dos criterios: el tipo de elemento devuelto por la consulta y el uso de funciones agregados. En pocas palabras, cuando se toma en consideración el tipo de elemento devuelto por la consulta se tiene que puede devolver un valor puntual o un conjunto de tuplas. En el contexto de las consultas probabilísticas esto puede tener implicaciones: para las consultas que buscan un valor los autores definen cotas superiores e inferiores que definen intervalos en los que los valores de la función deben estar, con una probabilidad acumulada de 1.

La segunda clasificación toma en cuenta si existen agregados o no. La presencia de agregados puede afectar como se verá la evaluación de consultas incompletas.

Además de estos dos trabajos se han examinado otros 3. Sin embargo, su aporte y utilidad para el presente de trabajo de investigación es menor. Chen, Chen y Xie proponen en "Cleaning Uncertain Data with Quality Guarentees" (2008) [3], una métrica para cuantificar la ambiguedad de una respuesta de consulta bajo semánticas de mundo posibles. Sobre esta base, se podría construir un mecanismo para "limpiar" la base de datos.

El trabajo "On databases with Incomplete Information" de Lipski (1981) [4] es muy citado por otros investigadores en el área de consultas incerteras y sin duda alguna constituye un hito muy importante en ésta area. Sin embargo, en dicho trabajo se busca tratar el tema con un formalismo matemático que va más allá de los alcances de este proyecto.

Por otro lado, Kang y Kim proponen en "Query type classification for web document retrieval"

(2003) [5] un mecanismo de clasificación de consultas para extracción de información del WEB. Sin embargo, dicha clasificación corresponde con el tipo de operación que se desea: úbicar algo por tópico, ubicar un homepage o ubicar un servicio. Dicha clasificación no está enmarcada dentro del presente trabajo de investigación y por ende tiene poca utilidad.

Por último, Rocquenco, Segoufin y Viano presentan en "Representing and querying XML with incomplete information" (2001) [6] un modelo para hacer consultas incompletas sobre XML. Existe cierto paralelismo con lo que se quiere hacer en este trabajo de investigación: realizar una segunda extracción de información cuando no se pueda contestar una consulta con lo que está en la ontología. Sin embargo, dicho trabajo no fue de utilidad para la clasificación de las consultas.

Capítulo 3

Análisis de incompletitudes en consulta y propuesta para determinar las fuentes de incompletitud

Capítulo 4

Diseño

Diseno.

Capítulo 5

Detalles de implementación

5.1. Consideraciones generales

5.1.1. Lenguaje de programación

El lenguaje de programación utilizado para la implementación de las seis metaheurísticas fue Java, compilado usando g++ en su versión 1.6.0_20.

Capítulo 6

Pruebas

Descripción del conjunto de pruebas ¡Hablar de los dominios: se trabajó sobre designaciones, etc¿
Duda: esto es una especie de marco metodológico?

6.1. Selección de Documentos de Prueba

La selección de los documentos se realizó sobre el conjunto de las actas de los Consejos Directivos y Académicos de la Universidad Simón Bolívar. Con el objetivo de probar apropiadamente el Extractor, se decidió trabajar con las actas desde el año 1998 hasta el año 2012. Esto permite que las pruebas se hagan sobre un conjunto de documentos semiestructurado, pero que puede mostrar variabilidad en su redacción y estructura en los años.

Del conjunto de actas en el intervalo de tiempo especificado, se hizo una preselección de los documentos para utilizar las actas de los Consejos Extraordinarios. Adicionalmente, según cada dominio de aplicación se extrajeron las actas que sí contienen información sobre los dominios. Es decir, se utilizaron los documentos que contienen información sobre los dominios elegidos.

Para probar el preprocesador, se tomó un conjunto de archivos seleccionados aleatoriamente con uniformidad sobre los años. La prueba consistió en ver si el fragmento de texto extraído por el preprocesador coincidía exactamente, estaba contenido, contenia o era completamente disjunto con el segmento de texto que debería ser extraído. Esta prueba se hizo manualmente.

6.2. Pruebas de Extractor Focalizado

Para probar el extractor focalizado, se tomó un conjunto de prueba seleccionado aleatoriamente con uniformidad sobre los años. De cada archivo seleccionado, se tomaban hasta 3 designaciones (falta explicalro para los otros dominios), que se utilizaban para realizar pruebas. Una prueba consite en hacer una búsqueda focalizada sobre cada uno de los campos de una designación: se asume que se tienen algunos

valores de esa designación (un contexto de extracción) y se buscan los campos faltantes.

Estas pruebas se hicieron variando 2 parámetros. En primera instancia, se varió sobre el minimum hit measure (varió entre 0.25;0.5;0.75;1) que mide la proporción de los campos con valores conocidos que aparecen en una unidad de extracción (designación). Luego para cada hit measure se varió sobre la probabilidad de que cada uno de los campos fuese desconocido. Esto es, para cada hit measure al hacer una prueba se determinaba aleatoriamente cuantos campos tenían valores conocidos. Esto para poder simular los casos de la vida real en los que no se tienen todos los campos de una designación menos el conocido. Las probabilidades de no tener un valor puntual variaron entre 0.1;0.25;0.75.

Capítulo 7

Resultados

Tabla 7.1: Resultados de la evaluación del Preprocesador de Textos

Dominio	Correctos	Incorrectos
Designaciones	.953	0.046
Designaciones	.953	0.046

Tabla 7.2: Resultados detallados la evaluación del Preprocesador de Textos

Dominio	**Aprobados**		**No aprobados**	
	Correctos	Con texto de más	Incompletos	Incorrectos
Designaciones	87.69 %	7.69 %	3.07 %	1.54 %

Tabla 7.3: Resultados detallados la evaluación del Preprocesador de Textos: Designaciones

Dominio	**Aprobados**		**No aprobados**	
	Correctos	Con texto de más	Incompletos	Incorrectos
Designaciones	87.69 %	7.69 %	3.07 %	1.54 %
	Aprobados: **95.39**		No aprobados: **4.61 %**	

El conjunto de prueba es de tama no 73, un tercio de la población.

Por que el extractor falla con designaciones? (Observaciones hechas al hacer las pruebas que pueden ser útiles en el análisis de datos)

1. Las unidades de informacion no son precisas. A veces hay multiples designaciones en una misma linea.

2. Hay multiples designaciones que coinciden en un mismo día para una misma persona.

3. Hay ratificaciones, correcciones, posteriores a la designacoines que modifican el resultado que uno pensaria correcto.

4. Hay errores de tipeo por parte de la secretaria.

5. Hay casos "patologicos"que es imposible generalizar con expresiones regulares que no sean hechas a la medida.

6. Hay casos en los que los campos tienden a tener muchos valores null y al no encontrar la respuesta en el

Tabla 7.4: Resultados detallados la evaluación del Preprocesador de Textos: Escalafón

Dominio	**Aprobados**		**No aprobados**	
	Correctos	Con texto de más	Incompletos	Incorrectos
Designaciones	80.51 %	12.98 %	6.4 %	0 %
	Aprobados: **93.6**		No aprobados: **6.4 %**	

El conjunto de prueba es de tama no 77, un tercio de la población.

Tabla 7.5: Resultados detallados la evaluación del Preprocesador de Textos: Jurados de Ascenso

Dominio	**Aprobados**		**No aprobados**	
	Correctos	Con texto de más	Incompletos	Incorrectos
Designaciones	77.55 %	16.32 %	0 %	6.12 %
	Aprobados: **93.88**		No aprobados: **6.12 %**	

El conjunto de prueba es de tama no 49, un tercio de la población.

mejor hit, se procede al segundo. Arreglar esto en el extractor? Hay muchos casos en los que un segundo match ayuda.

Tabla 7.6: Resultados de la evaluación del Extractor Focalizado - Dominio: Designaciones. UnitHit Measure mínimo:.75

Campo	Prob. Campo Faltante	**P. Aprobados**		**P. No aprobados**	
		Correctos	Con texto de más	Incompletos	Incorrectos
EsAsignado.calificacion	0	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.10	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.25	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.50	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
EsAsignado.fechaAsignacion	0	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.10	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.25	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.50	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
EsAsignado.fechaFinal	0	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.10	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.25	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.50	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
EsAsignado.motivo	0	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.10	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.25	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.50	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
Profesor.Nombre	0	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.10	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.25	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	
	0.50	87.69 %	7.69 %	3.07 %	1.54 %
		Aprobados: **95.38 %**		No aprobados: **4.61 %**	

Prob. Campo Faltante es la probabilidad de que no se tenga el valor uno de los campos que se utilizan para hacer extracción focalizada.

Conclusiones y recomendaciones

Destacar:

1. íngeniería del documento'. 2. Dependencia de un extractor más inteligente.

Bibliografía

[1] Aggarwal and Fellow, "A survey of uncertain data algorithms and applications", *IEEE*, vol. 21, pp. 609–623, 2009.

[2] D. KALASHNIKOV R. CHENG and S.PRABHAKAR, "Evaluating probabilistic queries over imprecise data", *Proc ACM SIGMOD*, 2003.

[3] J. CHEN R. CHENG and Xie X., "Cleaning uncertain data with quality guarantees", *Proceedings of the VLDB Endowmen*, vol. 1,1, pp. 722–735, 2008.

[4] Lipski Witold, "On databases with incomplete information", *Journal of ACM*, vol. 8,1, 1981.

[5] In-ho Kang and Kim GilChang, "Query type classification for web document retrieval", *Proc Sigir '03*, 2003.

[6] Segoufin Rocquenco and Viano, "Representing and querying xml with incomplete information", *PODS'01*, 2001.