ANALISA TUJUAN BEROBAT PASIEN BERDASARKAN JENIS PENYAKIT (ICD) MENGGUNAKAN TEKNIK DATA MINING ASSOCIATION RULES DENGAN ALGORITMA CLUSTERING

EDY KURNIAWAN
Email: kurniawardana@gmail.com

Fakultas Teknik
Universitas Muhammadiyah Ponorogo

ABSTRAK

Rekam medis atau biasa di sandikan dalam bidang kesehatan adalah ICD (International Calssification Diseases) merupakan rekaman dari riwayat pasien yang melakukan pengobatan di rumah sakit maupun balai pengobatan lainnya. Perkembangan ICD saat ini sudah mencapai level ICD-10. Hasil dari rekam medis tersebut saat ini meripakan tumpukan data yang belum diolah. Tumpukan data yang ada (rekam medis) saat ini difungsikan sebagai rujukan untuk mengambi tindakan terhadap pasien apabila di kemudian hari mengalami gangguan kesehatannya. Rekam medis saat ini sudah menjadi kegiatan rutinitas Rumah sakit ataupun puskesmas dan juga balai pengobatan lainnya yang mempunyai ijin praktek dari ikatan dokter indonesia. Dari rutinitas ini pada dasarnya bisa diambil informasi yang sangat besar dari hasil rekam medis yang selama ini hanya sebagai tumupukan data yang selalu bertambah setiap harinya. Data mining adalah solusi untuk mencari informasi yang terkandung dalam aktifitas rekam medis tersebut. Banyak informasi yang bisa diambil dari tumpukan data ICD di sebuah rumah sakit. Dari hasil pengolahan rekam medis tersebut bisa di cari pola penyakit yang di derita seseorang maupun di dasarkan dari wilayah asal pasien. Aplikasi Weka merupakan pembantu dalam penggalian tumpukan data yang saat ini hanya menjadi arsip saja. Rumah sakit Aisiyah merupakan subjek yang digunakan peneliti untuk melakukan pengolahan data tersebut dan kemudian untuk dicarikan pola dari penyakit pasien berdasarkan pada wilayah, jenis kelamin, dan klasifikasi umur.

Kata kunci: Rekam medis, ICD, dan rumah sakit.

PENDAHULUAN

Bahasa medis yang biasa dilakukan oleh dokter dalam melakukan diagnosa kemudian memberikan tindakan atas penyakit yang diderita pasien berupa bahasa kedokteran (rekam medik) yang selanjutnya hasil tersebut di kodekan oleh seorang ahli rekam medis menjadi kode-kode ICD. Kode ini adalah bahasa standart yang bisa digunakan oleh semua dokter meskipun bukan dokter spesialis untuk membacanya sesuai dengan aturan-aturan yang berlaku pada kode tersebut (ICD). ICD-10: "International Calssification diseases" dirancang untuk menyamakan perbandingan secara internasional, pengolahan pengelompokan klasifikasi, serta menyajikan statistik mortalitas dan morbiditas. Analisa situasi kesehatan secara umum dan pengelompokan penyakit, pemantauan kejadian dan masalah kesehatan dalam suatu kelompok yang berkaitan dengan variabel, seperti karakteristik dan individu yang terkena penyakit, sumber daya manusia, kualitas dan pedoman layanan

Berdasarkan model konsep analisis ditunjukkan solusi praktis dalam aturan pengkodean ICD. Arah masa depan dari penelitian ini adalah menggabungkan model sematik yang komprehensif dalam proses formalisasi, pengembangan variabelnya untuk mendukung pelengkapan

domain dan pemetaan indeks dari terminologi klinis yang ada.

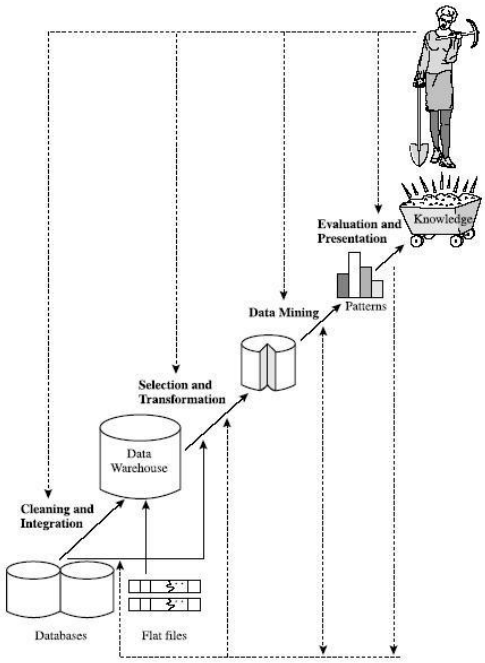
Dengan berkembangnya evidence based medicine dimana pelayanan medis yang berbasis data sangatlah diperlukan maka data dan informasi pelayanan medis yang berkualitas terintegrasi dengan baik dan benar sumber utamanya adalah data klinis dari rekam medis. Data klinis yang bersumber dari rekam medis semakin penting dengan berkembangnya rekam medis elektronik, dimana setiap entry data secara langsung menjadi masukan (input) dari sistem/manajemen informasi kesehatan. Dalam penelitian ini diusulkan dalam terminology hasil medical record mejadi kode ICD-10 menggunakan pengolahan menggunakan teknik data mining dengan algoritma clustering.

TINJAUAN PUSTAKA

Dengan meningkatnya transaksi yang disimpan dengan sistem basis data sekarang ini, maka dibutuhkan proses untuk menangani data tersebut. Proses untuk menangani data tersebut dikenal dengan *Knowledge Discovery in Databases (KDD)*. *Knowledge Discovery in Databases (KDD)* merupakan proses *nontrivial* dalam mengekstraksi data yang implisit, yang belum diketahui sebelumnya, dan berpotensi menjadi informasi yang berguna (Fayyad, Piatetsky-Shapiro, dan Smyth, 1996). *Nontrivial* karena beberapa pencarian atau inferensi yang dilibatkan bukan merupakan hasil komputasi secara langsung terhadap kuantitas yang telah didefinisikan sebelumnya, seperti komputasi nilai rata-rata sekumpulan bilangan. Pola yang ditemukan harus valid terhadap data baru pada suatu tingkat kepastian tertentu. Pola-pola tersebut harus dapat menjadi suatu deskripsi atau gambaran tentang suatu pengetahuan yang secara potensial berguna dan menguntungkan bagi pengguna atau tugas tertentu. Akhirnya, pola-pola tersebut juga harus dapat dipahami dan dimengerti, walaupun terdapat kemungkinan tidak dapat secara langsung dan harus melewati beberapa proses dahulu. Pada aplikasinya, sebenarnya *data mining* merupakan bagian dari proses *KDD.* Sebagai komponen dalam *KDD, data mining* terutama berkaitan dengan ekstraksi dan penghitungan pola-pola dari data yang ditelaah

Data Mining

Data mining adalah suatu proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database dalam jumlah besar. *Machine learning* adalah suatu area dalam *artificial intelligence(AI)* atau kecerdasan buatan yang berhubungan dengan pengembangan teknik-teknik yang bisa diprogramkan dan belajar dari data masa lalu.



Gambar 1. Alur Proses *KDD* [1]

Dalam Alur proses KDD pada gambar 1 dapat diuraikan sebagai berikut :

- *Data Cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi) [1].
- *Data Integration* merupakan proses menghubungkan antar data yang saling terkait untuk dapat memperkaya informasi yang dihasilkan. Hasil integrasi data sering diwujudkan dalam sebuah data warehouse karena dengan data warehouse, data dikonsolidasikan dengan struktur khusus yang efisien. Selain itu data warehouse juga memungkinkan tipe analisa seperti OLAP.
- *Data Selection* menciptakan himpunan data target, pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan (discovery) akan dilakukan analisa dan menghasilkan informasi yang berharga.
- *Data Transformation* merupakan proses transformasi data untuk menentukan kualitas dari hasil data mining, sehingga data diubah menjadi bentuk sesuai untuk di-Mining [1].
- *Data Mining* merupakan proses penggalian informasi Ada beberapa teknik data mining yang sudah umum dipakai.

Pattern Evaluation and Knowledge Presentation Dalam tahap ini hasil dari teknik data mining berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Presentasi pola yang ditemukan untuk menghasilkan aksi tahap terakhir dari proses data mining adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisa yang didapat.

Data Warehouse

Data warehouse didefinisikan dalam berbagai cara, sehingga sulit untuk ditentukan sebagai definisi yang terbaik. Singkat kata, data warehouse mengacu pada database yang dikelola secara terpisah dari database operasional dari sebuah organisasi. Data warehouse system memperbolehkan pengintegrasian dari beberapa aplikasi yang digunakan.

Menurut William H. Inmon, seorang arsitek pembangunan data warehouse system, “Suatu data warehouse adalah koleksi data yang subject-oriented, integrated, time-variant, dan nonvolatile dalam mendukung proses pengambilan keputusan oleh managemen.” Ungkapan

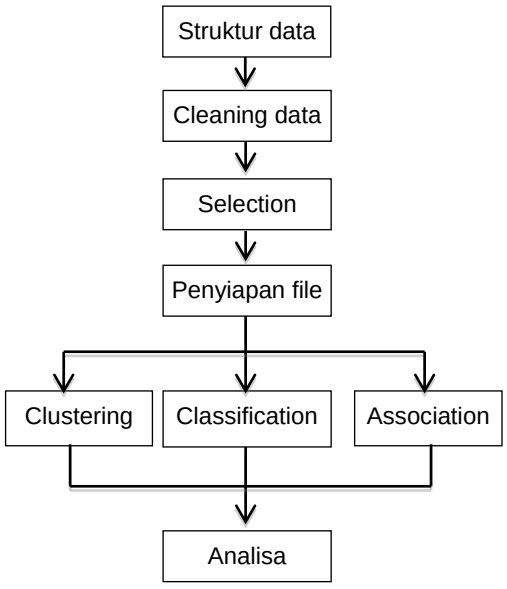
definisi itu setidaknya telah membedakan antara data warehouse dengan data repository lain seperti relational database system, transaction processing system, dan file system.

- Subject-oriented : Data warehouse di organisasikan pada beberapa subject, seperti customer, supplier, product, dan sales. Dengan tidak berkonsentrasi pada transaksi harian, data warehouse berfokus pada pemodelan dan analisa data untuk mendukung pengambilan keputusan. Singkatnya, data warehouse hanya terdiri dari data-data yang penting dan berpengaruh untuk mendukung pengambilan keputusan.
- Integrated : Data warehouse dibangun dari berbagai data sumber yang heterogen, seperti relational database, flat file, dan rekord treansaksi on-line. Data cleaning dan data integration diterapkan untuk menjaga konsistensi pada penamaan conventions, encoding structures, attribute measures, dan sebagainya.
- Time-variant : Data disimpan untuk menyediakan informasi historikal (seperti : 5 – 10 tahun) yang terkunci berdasarkan waktu.
- Nonvolatile : Sebuah data warehouse secara fisik selalu terpisah dengan data operasional, sehingga data warehouse tidak membutuhkan mekanisme pengendalian transactional processing, recovery, dan concurency. Biasanya ini hanya menggunakan dua operasional yaitu mengupload data dan mengakses data.

Sebagai kesimpulan, sebuah data warehouse adalah suatu data yang konsisten yang disajikan untuk mendukung pengambilan keputusan strategis organisasi. Konstruksi dari data warehouse membutuhkan adanya *data cleaning*, *data integration*, dan *data consolidation*. Penggunaan data warehouse biasanya membutuhkan ketersediaan teknologi *decision support*. Hal ini memberikan akses bagi *"knowledge worker"* (seperti manager, analis, eksekutif) untuk menggunakan data warehouse secara cepat dan meyakinkan dalam menampilkan data, dan membuat keputusan berdasarkan informasi yang tersedia di warehouse

METODE PENELITIAN

Langkah-langkah yang digunakan untuk pengolahan data rekam medis dengan program Weka 3.6.6 adalah sebagai berikut :



Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara

manual. Perlu diingat bahwa kata *mining* sendiri berarti usaha untuk mendapatkan sedikit data berharga dari sejumlah besar data dasar. Karena itu *data mining* sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik dan basisdata. Beberapa teknik yang sering disebut-sebut dalam literatur *data mining* antara lain yaitu *association rule mining*,*clustering*, *klasifikasi*, *neural network,* dan lain-lain.

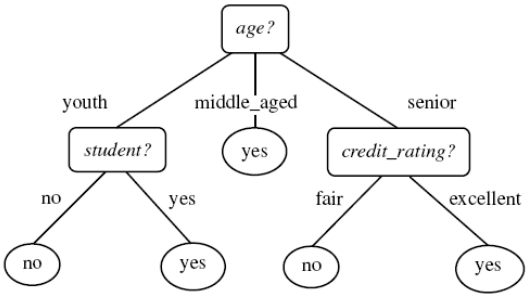
a. *Classification*

Suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah di definisikan.Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Aturan-aturan tersebut digunakan pada data-data baru untuk diklasifikasi. Teknik ini menggunkan *supervised induction*, yang memanfaatkan kumpulan pengujian dari record yang terklasifikasi untuk menentukan kelas-kelas tambahan. Salah satu contoh yang mudah dan populer adalah dengan Decision tree yaitu salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi. Decision tree adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. *Decision tree* adalah struktur *flowchart* yang menyerupai *tree* (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun merepresentasikan kelas atau distribusi kelas. Alur pada *decision tree* di telusuri dari simpul akar ke simpul daun yang memegang prediksi kelas untuk contoh tersebut. *Decision tree* mudah untuk dikonversi ke aturan klasifikasi (*classification rules*).

Ekstraksi pola pengelompokan atau pengklasifikasian sebuah himpunan obyek / data (training-set) ke dalam kelas (class) tertentu berdasarkan atribut-atributnya. Pola pengelompokan yang ditemukan akan menjadi model pengelompokan. Model digunakan untuk memprediksi kelompok data/obyek baru (test-set).

Klasifikasi data adalah suatu proses yang menemukan properti-properti yang sama pada sebuah himpunan obyek di dalarn sebuah basis data, dan mengklasifikasikannya ke dalam kelas-kelas yang berbeda menurut model klasifikasi yang ditetapkan. 'Untuk membentuk sebuah model klasifikasi, suatu sampel basis data 'E' diperlakukan sebagai training set, dimana setiap tupel terdiri dari himpunan yang sama yang memuat atribut yang beragam seperti tupel-tupel yang terdapat dalam suatu basis data yang besar 'W'. Setiap tupel diidentifkasikan dengan sebuah label atau identitas kelas. Tujuan dari klasifikasi ini adalah pertama-tama untuk menganalisa training data dan membentuk sebuah deskripsi yang akurat atau sebuah model untuk setiap kelas berdasarkan feature-feature yang tersedia di dalam data itu. Deskripsi dari masing-masing kelas itu nantinya akan digunakan untuk mengklasifikasikan data yang hendak di test dalam basis data 'W', atau untuk membangun suatu deskripsi yang lebih balk

untuk setiap kelas dalam basis data. Contoh untuk model ini adalah prediksi terhadap resiko pemberian kredit. Data terdiri



b. *Clustering*

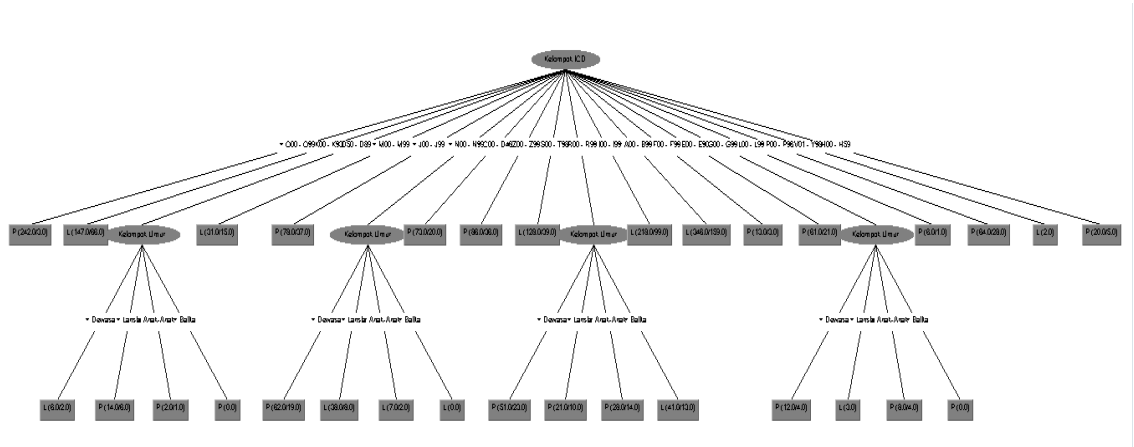
Clustering adalah proses pengelompokan sejumlah data/obyek kedalam kelompok-kelompok data (klaster) sehingga setiap klaster akan berisi data yang saling mirip. *Clustering* adalah salah satu teknik *unsupervised learning* dimana kita tidak perlu melatih metode tersebut atau dengan kata lain, tidak ada *fase learning.* Tujuan dari metode clustering adalah untuk mengelompokkan sejumlah data atau objek kedalam klaster sehingga setiap klaster akan terisi data yang semirip mungkin (Budi Santosa, 2007).

Ada dua jenis data *clustering* yang sering dipergunakan dalam proses pengelompokan data yaitu *hierarchical* (hirarki) data clustering dan *non-hierarchical* (non hirarki) data *clustering*. Ada dua macam teknik klaster yang cukup sering dipakai. Yang pertama adalah *k-means* (termasuk partitioning cluster) dan yang berikutnya adalah *hierarcichal clustering*

c. *Association*

Analisis asosiasi atau *association rule mining* adalah teknik data mining untuk menemukan aturan assosiatif antara suatu kombinasi item. Contoh dari aturan assosiatif dari analisa pembelian di suatu pasar swalayan adalah dapat diketahuinya berapa besar kemungkinan seorang pelanggan membeli roti bersamaan dengan susu. Dengan pengetahuan tersebut pemilik pasar swalayan dapat mengatur penempatan barangnya atau merancang kampanye pemasaran dengan memakai kupon diskon untuk kombinasi barang tertentu. Karena analisis asosiasi menjadi terkenal karena aplikasinya untuk menganalisa isi keranjang belanja di pasar swalayan, analisis asosiasi juga sering disebut dengan istilah *market basket analysis*

HASIL DAN PEMBAHASAN



=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: 4atributek_kelumur-weka.filters.unsupervised.attribute.Remove-R3
Instances: 1808
Attributes: 3
Gender
Kelompok Umur
Kelompok ICD
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

Kelompok ICD = O00 - O99: P (242.0/3.0)
Kelompok ICD = K00 - K93: L (147.0/66.0)
Kelompok ICD = D50 - D89
| Kelompok Umur = Dewasa: L (6.0/2.0)
| Kelompok Umur = Lansia: P (14.0/6.0)
| Kelompok Umur = Anak-Anak: P (2.0/1.0)
| Kelompok Umur = Balita: P (0.0)
Kelompok ICD = M00 - M99: L (31.0/15.0)
Kelompok ICD = J00 - J99: P (78.0/37.0)
Kelompok ICD = N00 - N99
| Kelompok Umur = Dewasa: P (62.0/19.0)
| Kelompok Umur = Lansia: L (38.0/8.0)
| Kelompok Umur = Anak-Anak: L (7.0/2.0)
| Kelompok Umur = Balita: L (0.0)
Kelompok ICD = C00 - D48: P (73.0/20.0)
Kelompok ICD = Z00 - Z99: P (86.0/36.0)
Kelompok ICD = S00 - T98: L (128.0/39.0)
Kelompok ICD = R00 - R99
| Kelompok Umur = Dewasa: P (51.0/23.0)
| Kelompok Umur = Lansia: P (21.0/10.0)
| Kelompok Umur = Anak-Anak: P (28.0/14.0)
| Kelompok Umur = Balita: L (41.0/13.0)
Kelompok ICD = I00 - I99: L (218.0/99.0)
Kelompok ICD = A00 - B99: L (346.0/159.0)
Kelompok ICD = F00 - F99: P (13.0/3.0)
Kelompok ICD = E00 - E90: P (61.0/21.0)
Kelompok ICD = G00 - G99
| Kelompok Umur = Dewasa: P (12.0/4.0)
| Kelompok Umur = Lansia: L (3.0)
| Kelompok Umur = Anak-Anak: P (8.0/4.0)
| Kelompok Umur = Balita: P (0.0)
Kelompok ICD = L00 - L99: P (6.0/1.0)
Kelompok ICD = P00 - P96: P (64.0/28.0)
Kelompok ICD = V01 - Y98: L (2.0)
Kelompok ICD = H00 - H59: P (20.0/5.0)

Number of Leaves : 31

Size of the tree : 36

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	1112	61.5044 %
Incorrectly Classified Instances	696	38.4956 %
Kappa statistic	0.2401	
Mean absolute error	0.4176	
Root mean squared error	0.4621	
Relative absolute error	84.6633 %	
Root relative squared error	93.0437 %	
Total Number of Instances	1808	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.552	0.305	0.695	0.552	0.615	0.673	P
	0.695	0.448	0.551	0.695	0.615	0.673	L
Weighted Avg.	0.615	0.368	0.632	0.615	0.615	0.673	

=== Confusion Matrix ===

a	b	<--	classified as
557	452	|	a = P
244	555	|	b = L

KESIMPULAN DAN SARAN

Dua ribu data yang di uji menggunakan weka 3.6.6 di dapat 1808 data yang mempunyai kelayakan uji sesuai variabel yang diambil (alamat, jenis kelamin, umur, kelompok ICD). Hal ini disebabkan karena pemurnian data uji dan dibuangnya data tidak falid yang nantinya *akan* menyebabkan noise dalam pembacaan program.

Dari *metode* algoritma yang digunakan Clasify menggunakan Decession Three J48 mempunyai hasil accurasi data 33.9049 %. Sedangkan pada clasify menggunakan decession three J48 yang mengacy pada clasifi kelompok ICD mempunyai titik accurasi 61.5044 %.

Hasil tersebut menguatkan bahwa dalam metode pengambil keputusan (menggunakan metode clasify decession three j 48) mempunyai perbedaan. Perbedaan ini di tentukan oleh faktor penentuan keputuan variabel yang berpengaruh pada data secunder. Jadi untuk mengambil keputusan pada data rekam medis mempunyai accurasi yang tinggi mengunakan decession three J48 dengan penentu adalah kelompok ICD.

DAFTAR PUSTAKA

Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze (2008), *Introduction to Information Retrieval*, Cambridge University press.

Dwi Cahyono, Junaidillah Fadlil, Suryo Sumpeno, Mochamad Hariadi (2008), "*Temu Kembali Informasi untuk Pembangkitan Basis* Pengetahuan *dari Teks Bebas yang Digunakan oleh Agen Percakapan Bahasa Alami*", Jurusan Teknik Elektro Institut Teknologi Sepuluh Nopember, Surabaya

Gorys Keraf, Dr. (1991), *Tata Bahasa Indonesia*, Volume XIV, Nusa Indah.

Guoqian Jiang, Jyotishman Pathak, Christoper G. Chute (2009), "*Formalizing ICD Coding Rules using Formal Concep Analysis*", Mayo Clinic Collage of Medicine, Rochester, USA.

Hermann Helbig (2006), Knowledge Representation and the Semantics of *Natural Language*, Springer-Verlag Berlin Heidelberg.

Ian H. Witten and Eibe Frank (2005), *" Data Mining Practical Machine Learning* Tools *and Techniques"*, Morgan Kaufmann Publishers is an imprint of Elsevier., San Francisco.

ICD-10 (International Classification of Diseases) [http://www.who.int/ classifications/icd/en/] Accesed Jan 2011

Jiawei Han and Micheline Kamber, *"Data Mining Concepts and Techniques",* Second Edition, 2006.

Stephane Meystre, Peter J. Haug (2005), "*Natural Language Processing to Extract Medical Problems from Electronic Clinical Document*", Departement of Medical Informatics, University of Utah school of Medicine, Salt lake City, UT, USA

Jisheng Liang, Thien Nguyen, Krzysztof Koperski, Giovanni Marchisio, Ontology-*Based Natural Language Query Processing for the Biological Domain*, Insightful Corporation.

Karim Sehaba, Vincent Courboulay and Pascal Estraillier (2003), *Interactive system by observation and analysis of behavior for children with autism*, Université de La Rochelle, France

Klas Burén, Andreas Ling (2005), *Natural Language Processing in a* Dialog *Based Assistant*, School of Mathematics and Systems Engineering Växjö University.

Lily Suryana Indradjaja, Stephane Bressan (2003), *Automatic Learning of Stemming Rules for the Indonesian Language*, PACLIC hal. 62-68.

Natalya F. Noy and Deborah L. McGuinness (2000), *Ontology*

Development 101: A Guide to Creating YourFirst Ontology, Stanford University, Stanford

Neil Matthew, Richard Stones (2005), *Beginning Databases with PostgreSQLFrom Novice to Professional, Second Edition,*Apress

Ozlem Uzuner, Jonathan Mailoa, Russell Ryan, Tawanda Sibanda (2010), "*Semantic Relations for Problem-Oriented Medical Records*", University at Albany, State University of New York, USA.

Pavel Berkhin, *"Survey of Clustering Data Mining Techniques"*, Accrue Software, Inc.

William J. Raynor, Jr (1999). *The International Dictionary of Artificial Intelligenc,* Glenlake Publishing Company, Ltd. Chicago, London, New Delhi Amacom.