ISSN: 2089-1121

INDUKSI DECISION TREE UNTUK PREDIKSI JUMLAH BURUH YANG BERHENTI BEKERJA

Arya Tandy Hermawan*), F.X. Ferdinandus*), David Boy Tonara*), dan Hartarto Junaedi)**
*) Program Pascasarjana Teknologi Informasi
Sekolah Tinggi Teknik Surabaya
**) Jurusan Sistem Informasi Bisnis
Sekolah Tinggi Teknik Surabaya
arya@stts.edu , ferdi@stts.edu , davidboy.tonara@gmail.com , aikawa@stts.edu

ABSTRAK

Satu masalah yang dihadapi oleh setiap bagian kepegawaian di perusahaan adalah *turnover* pekerja. Seringkali pekerja yang telah diterima permohonan kerjanya oleh perusahaan, mengundurkan diri sebelum pekerja tersebut memberikan kontribusi yang nyata kepada perusahaan. Pada umumnya minimal seorang pekerja membutuhkan waktu dua bulan untuk ditraining dalam menjalankan tugasnya, dua bulan untuk melakukan tugasnya tanpa pendampingan, dan dua bulan untuk melakukan training kepada juniornya agar dapat disebut telah memberikan kontribusi kepada perusahaan.

Penelitian ini akan membahas cara memodelkan solusiyang dapat mengatasi kesulitan-kesulitan yang dialami oleh bagian kepegawaian. Salah satunya adalah memprediksi lama kerja pekerja, yang seringkali menjadi titik lemah pada sebuah proses produksisehingga mengakibatkan hasil produksi tidak stabil dan akhirnya menyebabkan sejumlah masalah seperti rusaknya bahan baku, potensi keuntungan yang hilang, beban upah lembur yang meningkat, dan kesulitan dalam memprediksi pasar.

Sebuah alternatif solusi yang ditawarkan oleh teknologi informasi adalah ekstraksi pengetahuan dari data kepegawaian sehingga menghasilkan prediksi kebutuhan pekerja. Diharapkan dari pengetahuan tersebut, bagian kepegawaian dapat meramalkan kebutuhan pekerja tiap divisi pada setiap bulannya untuk menjaga kestabilan hasil produksi. Representasi pengetahuan yang ditawarkan adalah pemodelan dalam bentuk *decision tree*karena sifatnya yang paling mudah dipahami dan diterima oleh orang awam.

Kata kunci : Data Mining, Kepegawaian, Decision Tree

ABSTRACT

A problem faced by every general affairs in company is employee turnover. A lot of workers, who have been recruited by the company resigned before the worker give real contribution to the company. In general, a worker needs at least two months to take training for carrying out their duties, two months to do his/her job without assistance, and another two months to train his/her juniors to be considered as giving contribution to the company.

In this research will explain about how decision tree which formed by Knowledge Discovery from Database process, give way out to general affairs to

ISSN: 2089-1121

overcome their difficulties. The difficulties of general affairs are to predict the length of work for every worker often becomes a weakness in a production process. So this condition brings the company to an unstable state of production result which eventually leads to some problems like raw materials waste, potential profits loss, overtime expenses increasing, and market predicting difficulties.

An alternative resolution offered by information technology is knowledge discovery from database to produce prediction system of labor deficiency. With this system, general affairs can predict labor deficiency for each division for each month to maintain stable state of production. A knowledge representation which offered by this research is Decision Tree, because it's the easiest model to understand and accept by common people.

Keywords : Data Mining, Employee Affair, Decision Tree

1. PENDAHULUAN

Satu masalah yang dihadapi oleh setiap bagian kepegawaian di perusahaan adalah *turnover* pekerja. Seringkali pekerja yang telah diterima permohonan kerjanya oleh perusahaan, mengundurkan diri sebelum pekerja tersebut memberikan kontribusi yang nyata kepada perusahaan. Pada umumnya minimal seorang pekerja membutuhkan waktu dua bulan untuk ditraining dalam menjalankan tugasnya, dua bulan untuk melakukan tugasnya tanpa pendampingan, dan dua bulan untuk melakukan training kepada juniornya agar dapat disebut telah memberikan kontribusi kepada perusahaan.

Kesulitan dari bagian kepegawaian dalam memprediksi lama kerja pekerja inilah yang seringkali menjadi titik lemah pada sebuah proses produksi. Sehingga yang terjadi adalah hasil produksi tidak stabil yang akhirnya menyebabkan terjadi masalah-masalah seperti rusaknya bahan baku, potensi keuntungan yang hilang, beban upah lembur yang meningkat, dan kesulitan dalam memprediksi pasar.

Tujuan dari penelitian ini adalah menerapkansebuah sistem prediksi yang didasari ekstraksi pengetahuan pada data kepegawaian. Diharapkan sistem yang dibangun dapat memberikan prediksi jumlah pekerja yang dibutuhkan pada kurun waktu tertentu. Dengan demikian, sistem prediksi ini akan mengubah dasar jumlah penerimaan pekerja dari yang sebelumnya berdasarkan jumlah pekerja yang keluar di bulan tersebut menjadi jumlah pekerja yang diprediksi keluar di tiga bulan mendatang.

2. TINJAUAN PUSTAKA

Istilah *data mining* mengacu kepada kegiatan melakukan ekstraksi pengetahuan dari data dengan ukuran yang sangat besar. Istilah ini seharusnya kurang tepat jika mengacu pada kata lain yang serupa misalnya menambang emas atau menambang minyak. Istilah yang lebih tepat bisa berupa "ekstraksi pengetahuan dari data", yang lebih tidak familiar dibandingkan istilah "data mining (penambangan data)".

Classification adalah adalah salah satu bentuk terapan data mining yang menghasilkan suatu model pengetahuan. Model pengetahuan ini akan digunakan untuk menentukan jenis data baru atau memprediksi kecenderungan data yang akan datang. *Classification* memiliki tiga tahap pada prosesnya. Tahap pertama adalah *datapreprocessing* yang bertujuan untuk mempersiapkan data untuk proses *classification*. Data yang digunakan sebagai sumber pengetahuan harus telah ditentukan nilai keluarannya (classnya). Tahap berikutnya adalah membangun sebuah *classifier* (*knowledge base*) berdasarkan sekumpulan data yang telah diketahui nilai keluarannya

ISSN: 2089-1121

(classnya). Tahap awal ini seringkali disebut fase *training*, sedangkan tahap berikutnya adalah fase *testing* yang merupakan tahap penerapan *classifier* pada data yang belum memiliki nilai keluaran untuk ditentukan kemudian.

Jiawei Han dan Micheline Kamber menjelaskan bahwa *Data Preprocessing* dibagi menjadi 5 tahap yaitu *Data Selection* (Pemilihan Data), *Data Cleaning* (Pembersihan Data), *Data Integration and Trasformation* (Integrasi dan Transformasi Data), *Data Reduction* (Pengurangan Data), dan *Data Discretization* (Pemrosesan Data menjadi Data Diskrit).

Decision tree adalah representasi pengetahuan berupa pendekatan dari metode *divide and conquer*yaitu berupa rule yang ditampilkan dalam alur pertanyaan yang memiliki hanya jawaban ya atau tidak. Decision tree berawal dari sebuah *node* yang seringkali disebut *root node*, dari node tersebut akan berkembang menjadi tree yang sempurna melalui alternatif-alternatif nilai dari setiap node. Pada decision tree, input attribut direpresentasikan sebagai node, nilai dari input attribut direpresentasikan sebagai *branch*, sedangkan nilai keluaran (class) direpresentasikan sebagai leaf. Masalah utama pada decision tree terletak pada pemilihan atribut terbaik untuk dijadikan atribut pemisah pada urutan yang tepat. ID3 (Iterative Dichotomiser) adalah sebuah algoritma yang diperkenalkan oleh J. Ross Quinlan pada tahun 1986.

3. METODE PENELITIAN

Dalam pelaksanaan penelitian ini dibutuhkan beberapa tahap proses yang harus dilakukan. Tahapan-tahapan yang dilakukan adalah sebagai berikut:

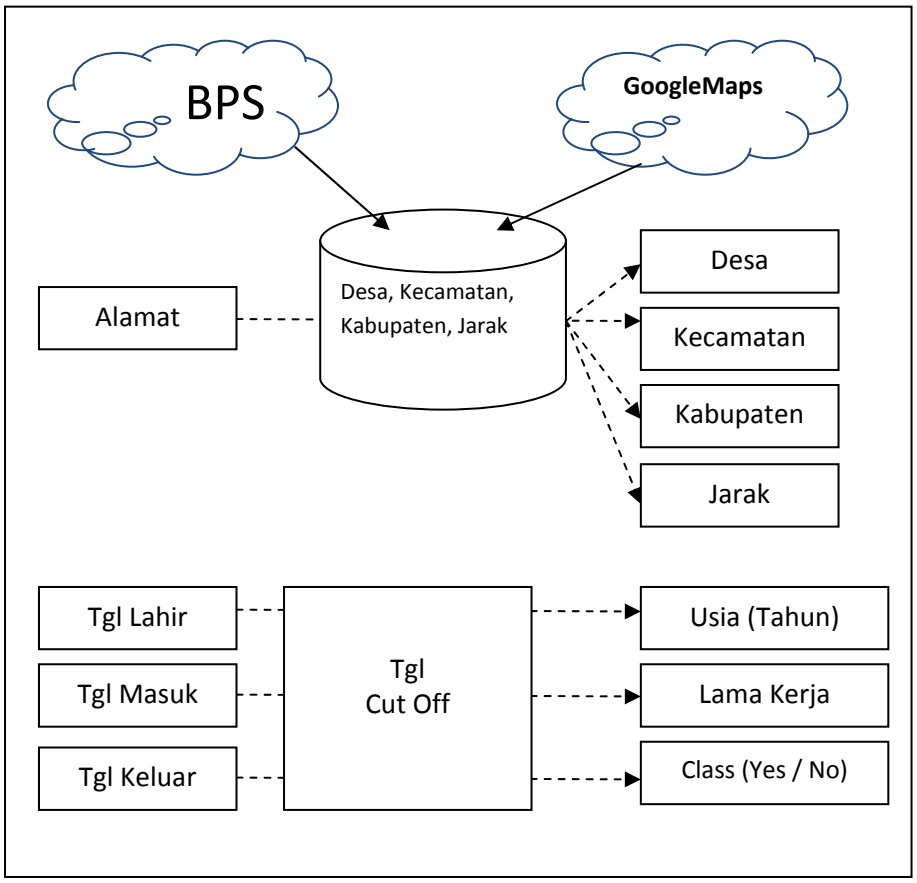
- Melakukan studi literatur mengenai metode yang akan digunakan, representasi pengetahuan yang akan digunakan, dan metode evaluasi akurasi hasil yang akan dilakukan.
- Melakukan data preparation terhadap data pekerja. Data pekerja yang diteliti adalah data pekerja dari PT Karunia Alam Segar, Manyar, Gresik yang mulai bekerja antara 1 Januari 2007 sampai 28 Februari 2011. Selain data pekerja, data yang digunakan adalah data absensi pekerja hingga Mei 2011. Data prediksi berdasarkan decision tree yang terbentuk nantinya akan dibandingkan dengan data kenyataan lapangan pekerja, untuk dihitung tingkat akurasi aktual berdasar data sebenarnya.
- Membangun sebuah decision tree sebagai representasi pengetahuan yang terekstrak dari data yang tersedia.
- Testing dengan menggunakan decision tree pada data sebenarnya.
- Evaluasi akurasi pemodelan decision tree berdasarkan data pekerja pada bulan berikutnya.

4. PREPROCESSING

Data Preprocessing adalah fase awal pada proses data mining yang bertujuan untuk mengubah data yang tidak ideal bagi proses data mining menjadi ideal untuk diproses. Data pada dunia nyata sangat rentan pada berbagai noda seperti, ketidak konsistenan data, kesalahan input data, kesalahan pengetikan, dan lain lain. Data pada dunia nyata berukuran terlalu besar sehingga proses data mining yang dilakukan kepada data tersebut akan menjadi tidak efektif. Selain itu seringkali data pada dunia nyata tidak cukup akurat untuk menggambarkan kondisi yang terwakili oleh data, karena itu dalam beberapa sisi perlu adanya sistem yang menimbulkan nilai-nilai tertentu untuk meningkatkan akurasi data.

ISSN: 2089-1121

Proses menimbulkan nilai-nilai baru untuk meningkatkan akurasi data ini pada fase data preprocessing disebut dengan Data Transformation. Untuk membentuk suatu sistem yang dapat melakukan prediksi terhadap setiap individu pekerja berkaitan dengan bertahan atau tidaknya pekerja tersebut dalam tiga bulan mendatang terdapat beberapa nilai yang perlu ditimbulkan untuk meningkatkan akurasi data. Nilai yang dianggap penting kaitannya dengan sistem prediksi ini adalah usia, lama kerja, desa, kecamatan, kabupaten, jarak dan jarak nominal.



Gambar 1. Tahap Transformasi Data

Dari gambar 1 dapat diketahui bahwa setidaknya ada dua proses yang terjadi pada fase data preprocessing yaitu transformasi data alamat dan transformasi data tanggal cut off. Transformasi data alamat membutuhkan inputan data dari Google dan BPS untuk menghasilkan attribute desa, kecamatan, kabupaten dan jarak. Sedangkan transformasi data tanggal cut off hanya melakukan perhitungan terhadap usia dan lama kerja pekerja serta menentukan label pekerja.

Transformasi data berikutnya yang dilakukan adalah dengan melakukan perubahan nilai jarak menjadi diskret untuk mengeliminasi attribute jarak, desa, kecamatan, dan kabupaten dari input attribute. Prosesini adalah menyatakan nilai jarak yang sebelumnya dinyatakan dalam bentuk attribute continous menjadi attribute nominal.

Pemilihan range nominal dari attribute didasarkan dari persebaran data pada data training Januari 2011. Berdasarkan persebaran data jarak pada data training Januari 2011, pembagian nilai dilakukan menjadi 5 nilai nominal. Kelima nilai nominal tersebut

ISSN: 2089-1121

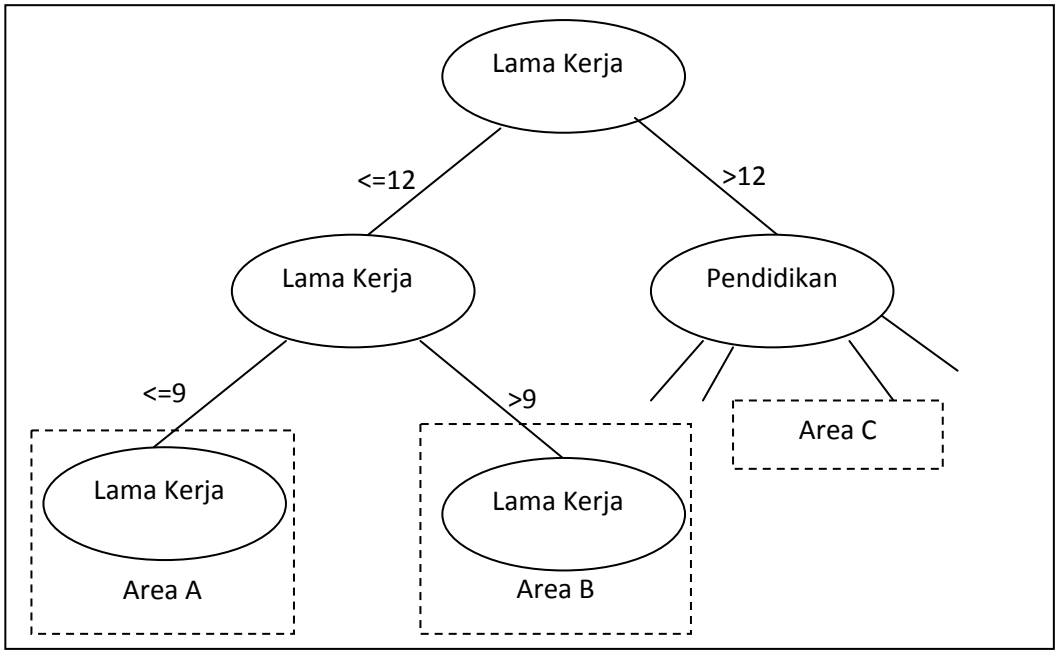
adalah untuk instance dengan nilai attribute jarak kurang dari atau sama dengan 10 km maka digolongkan ke dalam nilai nominal Ring 1, instance dengan nilai attribute jarak kurang dari atau sama dengan 30 km maka digolongkan ke dalam nilai nominal Ring 2, instance dengan nilai attribute jarak kurang dari atau sama dengan 90 km maka digolongkan ke dalam nilai nominal Ring 3, instance dengan nilai attribute jarak kurang dari atau sama dengan 180 km maka digolongkan ke dalam nilai nominal Ring 4, sedangkan instance dengan nilai attribute jarak lebih dari 180 km digolongkan ke dalam nilai nominal Ring 5.

5. INDUKSI DECISION TREE

Secara umum proses induksi decision tree terbagi ke dalam dua fase yaitu data training dan data testing. Fase training adalah fase penerapan algoritma (dalam hal ini ID3) ke dalam dataset yang siap dimining yang akhirnya menghasilkan representasi pengetahuan berupa decision tree. Fase testing adalah fase untuk menerapkan decision tree yang sudah terbentuk kepada data saat ini, sehingga dapat diketahui prediksi sistem berdasarkan pengetahuan yang telah ada kepada data secara individual maupun secara keseluruhan.

5.1 Fase Training

Tree yang dihasilkan berdasarkan data training Januari 2011 memiliki jumlah daun 298 buah, yang berarti memiliki 298 jalur untuk menuju class attribute dari root. Nilai akurasi decision tree yang dihasilkan oleh algoritma ini adalah 89.29%. Nilai akurasi yang digunakan di sini adalah nilai akurasi decision tree dengan menggunakan alternatif data testing Cross-Validation 10 Folds. Berdasarkan nilai-nilai penentu kesederhanaan dan akurasi di atas, transformasi data dengan menggunakan diskretisasi attribute jarak untuk menggantikan attribute desa, kecamatan, kabupaten, dan jarak adalah salah satu bagian dari tahap data preprocessing yang baik.



Gambar 2. Top Level Decision Tree Januari 2011

ISSN: 2089-1121

Berdasarkan gambar 2 dapat dilihat bahwa best split attribute yang pertama adalah lama kerja yang berarti attribute lama kerja paling menentukan apakah pekerja akan berhenti atau tidak dalam tiga bulan ke depan.Area A adalah area decision tree dengan split attribute awal jika Lama Kerja kurang atau sama dengan sembilan bulan. Area B adalah area decision tree dengan kriteria memiliki lama kerja lebih dari sembilan bulan dan kurang atau sama dengan dua belas bulan. Sedangkan Area C adalah area decision tree dengan kriteria lama kerja lebih dari dua belas bulan dan berikutnya akan dipisahkan berdasarkan input attribute pendidikan. Fase training pada penelitian ini akan menggunakan WEKA yang merupakan aplikasi penyedia layanan Data Mining hasil rancangan tim dari Universitas Waikato, Selandia Baru.

5.2 Fase Testing

Fase ini adalah fase yang bertujuan untuk mengaplikasikan decision tree sebagai representasi pengetahuan untuk memprediksi data yang belum memiliki label class.Setelah testing set ditentukan maka WEKA akan membentuk decision tree berdasarkan dataset training yang telah ditentukan pada langkah terdahulu, lalu menerapkan decision tree tersebut pada dataset testing yang baru saja ditentukan. Proses tersebut akan berjalan dan akurasi decision tree pada data real dapat diperoleh.

Hasil akurasi dari fase testing yang didapatkan dari fase training menggunakan dataset pekerja PT Karunia Alam Segar Januari 2011 dan fase testing menggunakan dataset pekerjanyauntuk bulan Maret 2011 adalah 90,29%. Hasil tersebut didapatkan dari 381 data diprediksi salah dan 3541 data diprediksi dengan tepat.

6. EVALUASI AKURASI

Evaluasi dilakukan dengan beberapa kasus yang berbeda, masing-masing adalah berdasarkan data training yang berbeda. Pada decision tree yang dibentuk berdasarkan data training Januari 2011 dan digunakan untuk memprediksi class attribute dari data April 2011, 3561 data dari 3940 data secara tepat diklasifikasikan. Jumlah data pekerja yang secara benar diklasifikasikan sebagai berhenti bekerja paling lambat tiga bulan setelah tanggal cut off adalah 101 pekerja, sedangkan 3450 data secara benar diklasifikasikan sebagai bertahan di pekerjaannya setidaknya tiga bulan setelah tanggal cut off yaitu 1 Januari 2011. Sehingga untuk decision tree dengan data training Januari 2011 yang digunakan untuk memprediksi data April 2011 adalah3561 / 3940 * 100% = 90.38 %.

Pada decision tree yang dibentuk berdasarkan data training Februari 2011 dan digunakan untuk memprediksi class attribute dari data Mei 2011, 3648 data dari 3810 data secara tepat diklasifikasikan. Jumlah data pekerja yang secara benar diklasifikasikan sebagai berhenti bekerja paling lambat tiga bulan setelah tanggal cut off adalah 105 pekerja, sedangkan 3543 data secara benar diklasifikasikan sebagai bertahan di pekerjaannya setidaknya tiga bulan setelah tanggal cut off yaitu 1 Januari 2011. Sehingga untuk decision tree dengan data training Januari 2011 yang digunakan untuk memprediksi data April 2011 adalah 3648 / 3810 * 100% = 95.75 %.

Pada decision tree yang dibentuk berdasarkan data training Maret 2011 dan digunakan untuk memprediksi class attribute dari data Juni 2011, 3837 data dari 3954 data secara tepat diklasifikasikan. Jumlah data pekerja yang secara benar diklasifikasikan sebagai berhenti bekerja paling lambat tiga bulan setelah tanggal cut off adalah 103 pekerja, sedangkan 3724 data secara benar diklasifikasikan sebagai bertahan di pekerjaannya setidaknya tiga bulan setelah tanggal cut off yaitu 1 Januari 2011.

ISSN: 2089-1121

Sehingga untuk decision tree dengan data training Januari 2011 yang digunakan untuk memprediksi data April 2011 adalah 3837 / 3954 * 100% = 97.04 %.

Berdasarkan decision tree yang dihasilkan dari pasangan-pasangan data training dan data testing yang telah dilakukan, akurasi sistem prediksi kebutuhan pekerja dengan menggunakan decision tree dapat dilihat pada tabel 1.

Tabel 1. Akurasi Prediksi oleh Decision Tree

Data Training	Data Testing	Akurasi
Desember 2010	Maret 2011	90.29 %
Januari 2011	April 2011	90.38 %
Februari 2011	Mei 2011	95.75 %
Maret 2011	Juni 2011	97.04 %

Sehingga rata-rata akurasi prediksi dengan menggunakan decision tree berdasarkan akurasi dari empat decision tree yang terbentuk dari empat data training yang berbeda adalah 93.37%.

7. PENUTUP

Dari hasil penelitian yang dilakukan, maka dapat diambil beberapa kesimpulan sebagai berikut. Dalam satu kesatuan proses data mining, fase data preprocessing adalah fase yang memakan waktu paling besar serta membutuhkan usaha yang jauh lebih besar dibandingkan fase-fase lain dalam proses data mining seperti fase training dan fase testing.Decision Tree adalah salah satu representasi pengetahuan yang cukup baik untuk lingkup non IT, karena Decision Tree mudah dimengerti serta dapat diaplikasikan berdampingan dengan pengetahuan dan pengalaman user.

Sedangkan dari hasil penelitian yang dilakukan, dapat diberikan beberapa saran untuk proses sejenis juga yang akan dilakukan perusahaan lainnya. Implementasi KTP Digital akan sangat membantu dalam penelitian-penelitian berbasis data mining yang menggunakan input attribute alamat atau penelitian-penelitian sejenis.Penelitian pada bidang data mining khususnya untuk menangani tipe data continous masih sangat diperlukan menimbang bahwa pemisahan attribute data continous yang ada menjadi binary kurang representatif untuk data-data besar.

8. DAFTAR PUSTAKA

Cabena, Peter,et al.*Discovering Data Mining: From Concept toImplementation.* Prentice Hall, New Jersey, USA. 1998.

Han, Jiawei. *Data Mining: Concepts and Techniques 2nd Edition*. Morgan Kaufmann Publishers, Inc. San Francisco. 2005.

Mitchell,TomMichael. *Machine Learning (International Edition).* MacGraw-Hill, Singapore.1997.

Quinlan, Ross. *Induction of Decision Trees,* Machine Learning 1, hal.81-106. 1986.

Witten, Ian H., Frank, Eibe. *Data Mining: Practical Machine Learning Tools and Techniques 2nd Edition*. Morgan Kaufmann Publishers, Inc. San Francisco. 2005.

Ye, Nong (Ed). *The Hand Book of Data Mining*. Lawrence Erlbaum Publisher. Mahwah, New Jersey. 1993.