JURUSAN TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER

USULAN TUGAS AKHIR

1. IDENTITAS PENGUSUL

Nama : **Noor Fitria Azzahra**
NRP : **5108100025**
Dosen Wali : **Yudhi Purwananto, S.Kom, M.Kom**

2. JUDUL TUGAS AKHIR

"Implementasi Analisa Weblog untuk Identifikasi Ketegori Pengguna dengan Pendekatan *Fuzzy* pada Studi Kasus Server ITS"

3. PENDAHULUAN

Personalisasi merupakan salah satu metode *web mining* yang bertujuan untuk melakukan kostumisasi interaksi sebuah situs agar dapat memberikan informasi yang sesuai dengan kebutuhan pengguna. *Data usage* (data kunjungan) merupakan salah satu sumber data yang dapat digunakan untuk melakukan personalisasi web. Data tersebut mencatat kunjungan setiap halaman pada sebuah website, termasuk waktu akses, alamat IP, dan lain-lain [1]. Data kunjungan dapat memberikan informasi mengenai pengguna yang mengakses halaman sebuah website sehingga dapat diketahui perilaku pengguna saat mengakses website tersebut. Data pengguna ini direpresentasikan oleh *server log file.*

Teknik *data mining* yang disebut dengan *Web Usage Mining* (WUM) dapat digunakan untuk mengetahui kecenderungan pengguna suatu web yang diambil dari *server log file* [2]. Dalam WUM, pengguna dikelompokan berdasarkan kesamaan ketertarikan terhadap suatu halaman website. Dalam jurnal [3], dicontohkan pengelompokan pengunjung www.dragonballit.ig antara lain: pengguna yang tertarik pada informasi tentang karakter dalam film; pengguna yang tertarik pada cerita sejarah pada film atau gambar-gambar karakter; pengguna yang hanya tertarik pada tokoh utama dalam film; pengguna yang menyukai halaman hiburan (permainan atau video); dan pengguna yang hanya tertarik pada informasi umum mengenai film tersebut. Dengan adanya informasi ini, pengelola

web dapat melakukan kostumisasi atau personalisasi website sehingga dapat memberikan informasi yang dibutuhkan oleh pengguna.

Pengelompokan ini menggunakan algoritma yang ada pada *data mining* yaitu *Clustering*. Metode yang digunakan harus disesuaikan dengan jenis data dan tujuan pengelompokan. Berdasarkan informasinya, sebuah halaman web dapat dikelompokan ke dalam lebih dari satu kelompok sehingga teknik pengelompokan menggunakan metode *fuzzy* [4]. Dalam WUM, data *log file* yang digunakan memiliki hubungan antara satu data dengan data lain sehingga disebut dengan data relasional [5]. Berdasarkan sifat kasus dan data yang digunakan, maka digunakan algoritma CARD+ (*Competitive Agglomeration for Relational Data)* yang menerapkan konsep *fuzzy* untuk pengelompokan data relasional [3]. Keunggulan CARD+ adalah dapat mencari jarak/kemiripan yang tidak terbatas pada cara *Ecluidean* [8]. Selain itu, CARD+ dapat menghilangkan partisi yang *redundant* ketika kelompok-kelompok tersebut memiliki derajat *overlapping* yang tinggi (*very low inter-cluster distance*) [3].

4. RUMUSAN DAN BATASAN MASALAH

Rumusan masalah yang diangkat dalam Tugas Akhir ini dijelaskan sebagai berikut:

1. Bagaimana metode pengolahan data *file log* pada server ITS sehingga dapat digunakan sebagai fitur untuk mengidentifikasi kategori pengguna.
2. Bagaimana cara mengolah data fitur sehingga dapat digunakan untuk mengelompokkan pengguna web berdasarkan perilaku terhadap suatu halaman website.

Adapun batasan masalah dalam pengerjaan Tugas Akhir ini adalah:

1. Data *file log* yang digunakan berasal dari web server ITS.
2. Algoritma *fuzzy* yang digunakan adalah CARD+.

5. TUJUAN DAN MANFAAT TUGAS AKHIR

Tujuan pengerjaan Tugas Akhir ini adalah melakukan analisis *file log* sehingga dapat dilakukan pengelompokan pengguna web berdasarkan perilaku terhadap suatu halaman website tertentu.

Sedangkan manfaat pengerjaan Tugas Akhir ini adalah dapat mengetahui profil pengguna suatu web. Informasi tersebut nantinya dapat digunakan sebagai informasi pendukung dalam proses kostumisasi halaman web maupun personalisasi web dengan

menambahkan fitur pemberian rekomendasi halaman web yang sesuai sehingga pengguna dimudahkan untuk memperoleh informasi yang relevan.

6. RINGKASAN TUGAS AKHIR

Ringkasan Tugas Akhir berisi tentang gambaran umum pengerjaan Tugas Akhir. Ringkasan tersebut dibagi menjadi beberapa bagian, antara lain sumber dan tipe data, *web usage mining*, dan metode CARD+ yang akan digunakan untuk mengidentifikasi kategori pengguna.

Sumber dan Tipe Data

Data yang digunakan dalam *web user mining* (WUM) adalah *data usage* yang direpresentasikan oleh *server log file* yang meliputi *web server access logs* dan *application server logs*. Data log dikoleksi secara otomatis oleh aplikasi server. Data ini dapat digunakan untuk merepresentasikan perilaku penjelajahan pengunjung web. Setiap datanya berkorespondensi terhadap HTTP *request* dan satu *entry server access log*. Contoh data ini dapat dilihat pada Gambar 1. Pada *log entry* pertama, IP pengguna adalah "123.123.123.123". *Username* adalah "- -" dengan *timespan* "[26/Apr/2000:00:23:47-400]" dan mengakses halaman "http://search.netscape.com/Computers/Data_Formats/Document/Text/RTF' [6].

1	123.123.123.123 - - [26/Apr/2000:00:23:47 -0400] "GET /asctortf/ HTTP/1.0" 200 8130 "http://search.netscape.com/Computers/Data_Formats/Document/Text/RTF" "Mozilla/4.05 (Macintosh; I; PPC)"
2	123.123.123.123 - - [26/Apr/2000:00:23:48 -0400] "GET /pics/5star2000.gif HTTP/1.0" 200 4005 "http://www.jafsoft.com/asctortf/" "Mozilla/4.05 (Macintosh; I; PPC)"
3	123.123.123.123 - - [26/Apr/2000:00:23:50 -0400] "GET /pics/5star.gif HTTP/1.0" 200 1031 "http://www.jafsoft.com/asctortf/" "Mozilla/4.05 (Macintosh; I; PPC)"
4	123.123.123.123 - - [26/Apr/2000:00:23:51 -0400] "GET /pics/a2hlogo.jpg HTTP/1.0" 200 4282 "http://www.jafsoft.com/asctortf/" "Mozilla/4.05 (Macintosh; I; PPC)"
5	123.123.123.123 - - [26/Apr/2000:00:23:48 -0400] "GET /pics/wpaper.gif HTTP/1.0" 200 6248 "http://www.jafsoft.com/asctortf/" "Mozilla/4.05 (Macintosh; I; PPC)"

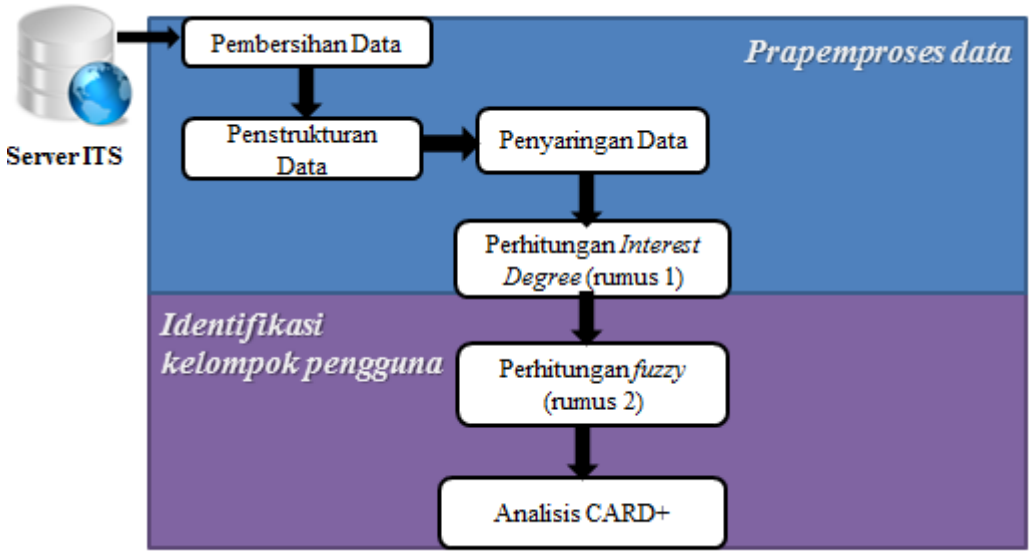
Gambar 1. Data server log

Web Usage Mining (WUM)

Web usage mining (WUM) merupakan sebuah metode untuk melakukan analisis pola dalam *clickstream*, data transaksi pengguna, maupun data terkait yang berhasil dikumpulkan sebagai hasil dari interaksi pengguna dengan website. Tujuannya adalah untuk mengetahui, memodelkan, dan menganalisis pola perilaku dan profil pengguna berdasarkan interaksinya terhadap website. Pola biasanya merepresentasikan koleksi

halaman, objek, atau link yang sering diakses atau digunakan oleh sekumpulan pengguna berdasarkan kebutuhan atau ketertarikan.

Proses WUM dibagi menjadi tiga, koleksi data dan preproses, deteksi pola, dan analisis pola. Dalam tahapan preproses, *log file* dibersihkan dan dipartisi menjadi himpunan transaksi pengguna yang merepresentasikan aktivitas setiap pengguna selama menjelajah sebuah situs. Dalam tahapan deteksi pola, metode statistika, database, maupun *mechine learning* digunakan untuk mengetahui pola tingkah laku pengguna yang tersembunyi. Pada tahapan akhir, pola tersebut diproses lebih jauh atau disaring sehingga dapat digunakan untuk membuat sistem rekomendasi, *visualization tools*, *web analytics* atau *report generation tools* [7]. Gambar 2 menjelaskan tentang alur proses WUM pada Tugas Akhir.



Gambar 2. Proses *web user mining* dengan menggunakan analisis CARD+

Tahap *Data Preprocessing*

Dalam tahap ini, dilakukan lima langkah, diilustrasikan pada Gambar 2. Pertama, pemilihan alamat website yang akan dianalisis dalam Tugas Akhir ini. Data web log pada server ITS jumlahnya cukup besar dan web yang diakses pun cukup banyak oleh karena itu diperlukan penentuan website yang akan digunakan sebagai objek penelitian. Langkah kedua adalah pembersihan data. Pada tahap ini, data *log file* dibersihkan dari atribut/fitur yang tidak dibutuhkan dalam proses pemodelan. Fitur *log file* yang dibutuhkan adalah data pengguna (IP/*email authentification*), waktu akses halaman, dan URL halaman tersebut. Langkah berikutnya adalah *data structuration*

(penstrukturan data). Pada tahap ini, *file log* yang telah dibersihkan dikumpulkan dan distrukturkan ke dalam bentuk *session. Session* adalah himpunan halaman yang diakses oleh pengguna yang sama dalam sebuah penjelajahan. Kumpulan pengguna yang mengakses website dilambangkan dengan $U = \{u_1, u_2, \dots, u_n\}$, n adalah jumlah pengguna website. Sedangkan *session* diidentifikasikan dengan $s_i = \langle u_i, t_i, p_i \rangle$ dimana u_i merupakan pengguna ke-i yang mengakses website, t_i adalah waktu akses user terhadap kumpulan halaman ke-i, dan p_i adalah himpunan halaman yang diakses dalam *session* ke-i. $p_i = \langle (p_{i1}, t_{i1}, pN_{i1}), (p_{i2}, t_{i2}, pN_{i2}), \dots., (p_{ij}, t_{ij}, pN_{ij}) \rangle$. j merepresentasikan indeks halaman. Dengan p_{ij} adalah halaman yang diakses selama *session* tersebut, t_{ij} adalah waktu akses halaman tersebut, dan N_{ij} adalah jumlah akses terhadap halaman tersebut selama *session* berlangsung. Langkah keempat adalah penyaringan data. Pada tahap ini, *session* dengan halaman yang paling jarang dikunjungi akan dihapus dari himpunan *session.* Langkah terakhir adalah menghitung derajat ketertarikan. Pada tahap ini, akan dihasilkan m x n matriks *behavior* (matriks B) = $[b_{ij}]$, n dan m mengindikasikan jumlah pengguna/*session* dan halaman yang dikunjungi. Dan b_{ij} merepresentasikan derajat ketertarikan pengguna *i* terhadap halaman web *j*. Kemudian dilakukan perhitungan derajat ketertarikan pada matriks B dengan rumus:

$$b_{ij} = \begin{cases} IG_{ij} & \text{Jika halaman } p_j \text{ diakses dalam session } s_i \\ 0 & \text{lainnya} \end{cases}$$

Dimana $IG_{ij} = f_{ij} . \frac{t_{ij}}{t_i}$ (1)

Keterangan:

- f_{ij} adalah frekuensi akses halaman ke-j selama session ke-i, $f_{ij} = N_{ij} / \sum_{k=1}^{n_i} N_{ik}$
- t_{ij} adalah waktu akses halaman ke-j pada session ke-i
- t_i adalah waktu total akses selama session ke-i

Pengelompokan *Log File* dengan CARD+

Clustering (pengelompokan) menjadi masalah penting dalam *data mining* dan *mechine learning*. Analisis pengelompokan adalah sebuah proses yang mempartisi sejumlah data objek menjadi beberapa *cluster* (kelompok). Setiap objek yang berada pada sebuah kelompok memiliki kesamaan dan objek yang berbeda kelompok dinyatakan memiliki perbedaan. Kebanyakan pendekatan pengelompokan ini berfokus pada data "*flat*" yang berarti bahwa setiap objek data direpresentasikan sebagai atribut vektor yang panjangnya

tetap. Pada kenyataannya, sebagian besar *dataset* memiliki struktur yang lebih rumit, termasuk adanya keterkaitakan antara satu data dengan data lain, seperti dokumen, kata-kata dalam korpus, *web pages,* atau *query*. Data inilah disebut dengan data relasional [7].

Dalam pengelompokan data, terdapat beberapa permasalahan sehingga memerlukan pemilihan metode khusus untuk menyelesaikannya. Misalnya, sebuah objek dapat menjadi anggota di dua kelompok sekaligus maupun jumlah kelompok optimal juga belum dapat diketahui. Permasalahan tersebut dapat diatasi dengan menggunakan algoritma *fuzzy Competitive Agglomeration Relational Data* (CARD+) yang dapat mengelompokan data ke dalam jumlah komponen optimal secara otomatis [8]. Selain itu, algoritma ini memiliki kelebihan dalam hal menghilangkan partisi yang *redundant* ketika kelompok-kelompok tersebut ada yang memiliki derajat *overlapping* yang tinggi (*very low inter-cluster distance*) [3].

Seperti algoritma *relational clustering* pada umumnya, CARD+ mempartisi data *relational object* misalnya data kuantitas relasi antara pasangan objek. Kuantitas relasi tersebut merepresentasikan derajat kemiripan antara dua vektor. Pada [3], CARD+ dilengkapi dengan *fuzzy similiarity measure* yang baru. Dimana setiap vektor dimodelkan sebagai kumpulan *fuzzy.*

Dalam referensi tersebut, matriks *behavior* B dikonversi menjadi matrix F=[μ_{ij}] yang merepresentasikan derajat ketertarikan dari pengguna terhadap suatu halaman dengan metode *fuzzy*. Derajat ketertarikan tersebut ditentukan dengan cara berikut:

$$\mu_{ij} = \begin{cases} 0 & \text{if } b_{ij} < ID_{min} \\ \frac{b_{ij} - ID_{min}}{ID_{max} - ID_{min}} & \text{if } b_{ij} \in [ID_{min}, ID_{max}] \\ 1 & \text{if } b_{ij} > ID_{max} \end{cases} \quad (2)$$

Dimana ID_{min} adalah *threshold* minimal sedangkan ID_{max} adalah *threshold* maksimal.

$$Sim_{(k,l)} = \frac{\sum_{j=1}^{m} min\{\mu_{k,j}, \mu_{l,j}\}}{\sum_{j=1}^{m} max\{\mu_{k,j}, \mu_{l,j}\}} \quad (3)$$

Dari matriks F, dibuat matriks *similarity* nxn $Sim = [sim_{ij}]$, n merupakan jumlah *session*/ pengguna. Inisialisasi matriks *Sim* dilakukan dengan persamaan 3. Dimana k dan l adalah indeks baris matriks F. Dari matriks Sim tersebut dibentuk matriks relasi $R = [r_{ij}]$. Matriks ini berukuran nxn yang memiliki nilai $R = 1 - Sim$.

Matriks-matriks tersebut digunakan untuk menentukan kelompok pengguna berdasarkan ketertarikannya terhadap halaman pada sebuah website. Pengelompokan ini menggunakan algoritma CARD+ yang dijelaskan pada algoritma berikut:

1. Inisialisasi:
   - jumlah *cluster* (kelompok) $C = C_{max}$ $(2 \leq C \leq n)$
   - *threshold* $\epsilon_1 = 5$
   - $k = 0;\ \beta = 0$
   - Inisialisasi matriks *fuzzy C-partition* $U^{(0)}$ matriks berukuran $C \times N$
   - $N_i = \sum_{j=1}^{n} u_{ij}$, $1 \leq i \leq C$, N_i merupakan nilai kardinal
   - vektor prototipe $V = \{v_1, v_2, .., v_c\}$
2. Ulangi sampai keanggotaan *fuzzy* stabil
   a. Menghitung vektor keanggotaan

      $$z_i = \frac{(u_{i1}, u_{i2}, .., u_{in})^t}{\sum_{j=1}^{n} u_{ij}}$$

   b. Menghitung $d_{ik} = (Rz_i)_k - z_i^t Rz_i/2$
   c. Jika $d_{ik} < 0$ untuk sembarang i dan k maka:
      i. Hitung $\Delta\beta = max_{ik} \left\{ -\frac{2d_{ik}}{\|z_i - e_k\|^2} \right\}$, e_k merupakan kolom ke-k dari matriks identitas
      ii. *Update* $d_{ik} = d_{ik} + (\Delta\beta/2) * \|z_i - e_k\|^2$ untuk $1 \leq i \leq C$ dan $1 \leq k \leq n$
      iii. *Update* $\beta = \beta + \Delta\beta$
   d. *Update* $\alpha(k) = \eta_0 \, e^{-k/\tau} \frac{\sum_{i=1}^{C} \sum_{j=1}^{n} (u_{ij})^2 d\,(b_j, \beta_i)}{\sum_{i=1}^{C} [\sum_{j=1}^{n} u_{ij}]}$ dimana η_0 adalah *exponential decay* diinisialisasi 5 dan τ adalah waktu konstan dengan nilai 10
   e. *Update* $U^{(k)} = U^{(FCM)} + U^{(Bias)}$

      Dimana $U^{(FCM)} = \frac{\frac{1}{d}(X, PV)}{\sum_{i=1}^{c} \frac{i}{d(x, PV_k)}}$ dan $U^{(Bias)} = \frac{\alpha}{d(x,\beta)(n - \bar{n})}$

      Dimana $\bar{n} = \frac{\sum_{i=1}^{c} \frac{1}{d}(X, pv_k) N_k}{\sum_{i=1}^{c} \frac{1}{d}(X, pv_k)}$
   f. Menghitung $N_i = \sum_{j=1}^{n} u_{ij}$

g. Jika ($N_i < \epsilon_1$) buang kelompok ke-i dan *update* C dan PV

h. k = k+1

3. Membuat kelompok $\chi_c = (b_i \in B | d_{ci} < \forall_c \neq k)$, 1 ≤ c ≤ C
4. Menghitung vektor prototipe $v_{cj} = \frac{\sum_{b_i \in \chi_c} b_{ij}}{|\chi_c|}$ j = 1,..., m
5. Menghitung jarak antar kelompok $D = [D_{ij}]_{i,j = 1...c}$
6. Menghitung nilai rata-rata ϵ dari $D_{ij}, i, j = 1.. , C\ i \neq j$
7. Jika (D_{ij} < ϵ untuk setiap i dan j) maka
   a. Gabungkan kelompok $v_i$dan v_j
   b. *Update* C
   c. *Update* prototipe kelompok $v_i, i = 1 ...., C$
   d. *Update* matriks partisi U
   e. *Kembali ke langkah 6*
8. Selain itu berhenti

Algortima CARD+ menghasilkan keluaran sebagai berikut:

- C kelompok prototipe yang direpresntasikan sebagai vektor $v_c = \{v_{c1}, v_{c2}, .., v_{cm}\}$ untuk c = 1, ... , C
- Matriks partisi *fuzzy* M = $[m_{ic}]_{i=1..n}^{c=1..c}$dimana m_{ic} merepresentasikan derajat keanggotaan dari perilaku pengguna pada vektor $b_i$terhadap kelompok ke-c.

7. METODOLOGI

Metodologi yang akan dilakukan dalam Tugas Akhir ini memiliki beberapa tahapan, diantaranya sebagai berikut :

1. Penyusunan proposal Tugas Akhir

   Tahap awal untuk memulai pengerjaan Tugas Akhir adalah penyusunan Proposal Tugas Akhir. Pada proposal ini, penulis mengajukan gagasan pendekatan metode *fuzzy* dengan algoritma CARD+ untuk mengelompokkan pengguna berdasarkan ketertarikan pengguna terhadap website tersebut.

2. Studi literatur

   Pada tahap ini dilakukan pencarian, pengumpulan, penyaringan, pembelajaran dan pemahaman literatur yang berhubungan dengan *web usage mining* dan *fuzzy*. Literatur yang digunakan dalam pengerjaan Tugas Akhir ini berasal dari internet berupa makalah ilmiah, tesis, artikel, materi kuliah, serta beberapa buku referensi.

3. Implementasi

   Implementasi merupakan tahap untuk membangun sistem tersebut.

4. Pengujian dan Evaluasi

   Pada tahap ini dilakukan uji coba terhadap sistem yang telah dibuat, mengamati kinerja sistem yang baru dibuat, serta mengidentifikasi kendala yang mungkin timbul.

5. Penyusunan buku Tugas Akhir

   Tahap terakhir merupakan penyusunan laporan yang memuat dokumentasi mengenai pembuatan serta hasil dari implementasi perangkat lunak yang telah dibuat.

8. JADWAL KEGIATAN TUGAS AKHIR

Jadwal kegiatan pengerjaan Tugas Akhir dijelaskan sebagai berikut;

Kegiatan	Bulan												
	Maret				April				Mei				Juni
Penyusunan proposal Tugas Akhir													
Studi literatur													
Standarisasi *file log*													
Pemodelan													
Evaluasi model pengelompokan													
Penyusunan laporan Tugas Akhir													

9. DAFTAR PUSTAKA

[1] Pierrakos D., Paliouras G. , Papatheodorou C. and Spyropoulos D. , "Web Usage Mining as a Tool for Personalization: A Survey," *User Modeling and User-Adapted Interaction*, vol. 13, pp. 311-372, 2003.

[2] B. Mobashe, "Web Usage Mining and Personalization," in *Practical Handbook of Internet Computing*, CRC Press, 2005, pp. 2-31.

[3] Castellano G, Fanelli A.M. and Torsello M.A., "NEWER: A System for NEuro-fuzzy WEb Recommendation," *Applied Soft Computing*, vol. 11, no. 1, pp. 793-806, January 2011.

[4] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Introduction to Data Mining, Minnesota: Addison-Wesley, 2006.

[5] Bo Long, Zhongfei Zhang and Philip S. Yu, Relational Data Clustering Models, Algorithms, and Application, CRC Press, 2010.

[6] J. Limited, “A Web Server Log File Sample Explained,” JafSoft Limited, 9 September 2005. [Online]. Available: http://www.jafsoft.com/searchengines/log_sample.html. [Accessed 2 March 2011].

[7] B. Liu, Web Data Mining Exploring Hyperlinks, Contents, and Usage Data Second Edition, London: Springer, 2011.

[8] Hichem Frigui and Raghu Krishnapuran, “Clustering by Competitive Agglomeration,” *Pattern Recognition*, vol. 30, no. 7, pp. 1109-1119, 1997.

[9] Olfa Nasraoui, Hichem Frigui, Raghu Krishnapuram and Anupam Joshi, “Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering,” *Artificial Intelligence Tools*, vol. 9, no. 4, pp. 509-526, 2000.