P-49

IMPLEMENTASI METODE SUPPORT VECTOR MECHINE DAN K-MEANS CLUSTERING DENGAN SUPPORT VECTOR MECHINE TEMA KARYA ILMIAH PADA STMIK WIDYA CIPTA DHARMA

IMPLEMENT THE SUPPORT VECTOR MECHINE METHOD AND K-MEANS CLUSTERING WITH SUPPORT VECTOR MECHINE ON THE THEME OF SCIENTIFIC WORK AT STMIK WIDYA CIPTA DHARMA

Andi Yusika Rangan[1*], M. Irwan Ukkas[2] , Siti Qomariah[3]
[1,2,3]STMIK Widya Cipta Dharma,Jl. M.Yamin No 25,Samarinda

[*]Email: Andi@wicia.ac.id

Diterima 20-10-2018	Diperbaiki 28-11-2018	Disetujui 20-12-2018

ABSTRAK

Tren topik penelitian berguna untuk membangunn roadmap penelitian sebuah institusi. Klasifikasi tema karya ilmiah sangan diperlukan untuk melakukan evaluasi kecenderuangaan topik penelitian sebuah institusi. Implementasi klasifikasi text mining untuk tema karya ilmiah mengunkan metode supprort Vector Mechine dan K-means Clustering dengan dukungan Support Vector Mechine (SVM). Hasil penelitian ini untuk mengukur akurasi dari masing masing metode yang digunakan. Metode Support Vector Mechine tingkat akurasinga sebesar 93.33% sedangkan metode K-means Clustering dengan dukungan Support Vector Mechine (SVM) tingkat akurasinya 99,33%.

Kata kunci:klasifikasi, SVM, K-Means_Clustering

ABSTRACT

Trends in research topics are useful for building an institutional research roadmap. The classification of themes of scientific works is very much needed to evaluate the vulnerability of research topics of an institution. The implementation of the text mining classification for the theme of scientific works uses the Vector Mechine supprort method and K-means Clustering with Support Vector Mechine (SVM) support. The results of this study are to measure the accuracy of each method used. Support Vector Mechine method at the level of accuracy is 93.33% while the K-means Clustering method with Support Vector Mechine (SVM) accuracy is 99.33%.

Keywords: klasifikasi, SVM, K-Means_Clustering

PENDAHULUAN

Banyaknya judul karya ilmiah pada sebuah institusi seperti kampus tergantung dari lamanya institusi berdiri dan banyaknya jumlah lulusan. Pengetahuan mengenai tren karya ilmiah yang dihasilkan oleh perguruan tinggi memberikan manfaat bagi pengembangan kurikulum dan roadmap penelitian bagi institusi. Berbagai karya ilmiah dari sivitas akademika seperti skripsi, laporan penelitian, laporan penulisan ilmiah, laporan kuliah kerja nyata yang tersimpan secara digital namun, pada umumnya fenomena ini tidak disertakan pengetahuan yang disarikan dari dokumen-dokumen elektronika tersebut.

Saat ini ilmu pengetahuan sudah sangat berkembang untuk mensarikan atau mempolakan judul-judul karya ilmiah dalam hal ini berbentuk text kedalam kelas-kelas tertentu. Menurut Gupta & Lehal [1] Metode text mining merupakan pengembangan dari data mining yang diterapkan untuk mengatasi masalah tersebut. Algoritma-algoritma dalam text mining di buat untuk mengenali data yang sifatnya semi terstuktur misalnya sinopsis, abstrak maupun isi dari dokumen-dokumen.

Kategorisasi teks dapat digunakan untuk melakuakn pengalian opini (*opinion mining)* dan analisa sentiment. Algoritma katagorisasi teks saat in banyak berkembang antara lain *support Vector Machine (*SVM), *Naïve Bayes*, C4.5, K-Nearest Neighbours (K-NN) dan lain-lain.

K-means clustering merupakan metode yang popular digunakan untuk mendapatkan deskripsi dari sekumpulan data dengan cara

mengungkapkan kecenderungan setiap individu data untuk berkelompok dengan individu data lainya. Pada penelitian ini K-means clustering digunaan untuk memperbaiki proses kalsifikasi data teks yaitu melakuan klasterisasi data agar tingkat akurasi model yang diusulkan menjadi lebih baik.

Algoritma *Support Vector Machine* (SVM) digunkan untuk klasifikasi penentuan jenis tema karya ilmiah dosen dan mahasiswa. SVM adalah metode yang banyak digunakan untuk klasifikasi data berupa teks dengan tingkat akurasi yang baik.

Penelitian penerapan algoritma K-means Clustering untuk Optimasi Klasifikasi Tema karya ilmiah dosen danMahasiswa mengunakan *Support Vector Machine (SVM)* Pada STMIK Widya Cipta Dharma bertujuan untuk proses penemuan pola pengelompokan berbagai topik tugas akhir mahasiswa yang bermanfaat menghasilkan informasi tren penelitian perguruan tinggi dari tahun ke tahun.

METODOLOGI

Menurut Witten [2], serangkaian proses mendapatkan pengetahuan atau pola dari kumpulan data disebut dengan data mining. Data mining memecahkan masalah dengan menganalisis data yang telah ada dalam database**.** Suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan.Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Klasifikasi menggunakan *supervised learning.*

Menurut Michael [3], *Text mining* adalah satu langkah dari analisis teks yang dilakukan secara otomatis oleh komputer untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen (Han & Kamber, 2006). Prosedur utama dalam metode ini terkait dengan menemukan kata-kata yang dapat mewakili isi dari dokumen untuk selanjutnya dilakukan analisis keterhubungan antar dokumen dengan menggunakan metode statistik tertentu seperti analisis kelompok, klasifikasi dan asosiasi. Tahapan dalam *text mining* secara umum adalah *tokenizing, filtering, stemming, tagging,* dan *analyzing. Tokenizing* merupakan tahapan untuk memisah-misahkan setiap kata (*token*) pada dokumen *input*. *Filtering* merupakan proses seleksi terhadap kata-kata yang dihasilkan dari proses *tokenizing*, dapat dilakukan dengan algoritma *stop list* maupun *word list*. Algoritma *stop list* akan membuang kata-kata yang tidak penting seperti kata ganti, kata keterangan, kata sambung, kata depan dan kata sandang. Sebaliknya, algoritma *word list* akan menyimpan kata-kata yang penting. Proses *stemming* kemudian dilakukan untuk mencari kata dasar dari setiap kata yang telah lolos proses *filtering*. Terdapat 4 varian algoritma untuk proses *stemming* ini, yaitu: (1) *Table lookup,* seluruh kata dasar disimpan dalam memori untuk selanjutnya dijadikan acuan dalam pemeriksaan dokumen *input*. Kelemahan metode ini adalah membutuhkan ruang penyimpanan yang besar; (2) *Successor variety,* setiap kata dalam dokumen *input* yang akan diperiksa dipecah secara bertahap menjadi awalan-awalan (prefiks). Untuk setiap awalan kemudian dicari kemungkinan bentuk lainnya (variasinya) didalam *corpus*, pencarian dihentikan jika jumlah temuan telah melampaui nilai batas tertentu; (3) *N-gram*, pemeriksaan setiap kata dalam dokumen *input* dilakukan dengan menerapkan konsep *clustering*. Setiap kata dicari nilai kedekatannya dengan kata-kata yang lain dan disimpan dalam sebuah matriks. Matriks tersebut kemudian dijadikan acuan untuk melakukan pengelompokan kata-kata; (4) *Affix removal*, untuk setiap kata pada dokumen *input* dihilangkan awalan dan akhirannya dengan mengacu kepada *action rules*. Proses *tagging* dilakukan untuk mencari bentuk awal dari setiap kata lampau. Setelah semua kata penting berhasil dikoleksi dari rangkaian proses tersebut, maka tahap berikutnya adalah *analyzing* yaitu menentukan keterhubungan antar dokumen dengan mengamati frekuensi kemunculan tiap kata yang ada pada tiap dokumen.

Menurut Turban [4], *K-Means Clustering* merupakan metode yang popular digunakan untuk mendapatkan dekripsi sekumpulan data dengan cara mengungkapakan kecenderungan setiap data untuk berkelompok dengan individu-individu data lainya. Kecenderungan penegelompokan tersebut didasarkan pada kemiripan karakteristik individu-individu data yang ada. Ide dasar dari teknik ini adalah menemukan pusat dari setiap kelompok data yang mungkin ada untuk kemudian mengelompokkan setiap data individul kedalam salah satu dari kelompok-kelompok tersebut berdasarkan jaraknya. K-Means Clustering Merupakan

metode clustering non hirarki, metode yang membagi data kedalam cluster-clusteryang memiliki karakteristik yang sama. Secara umum algoritma dasar dari K-Means Clustering adalah sebagai berikut :

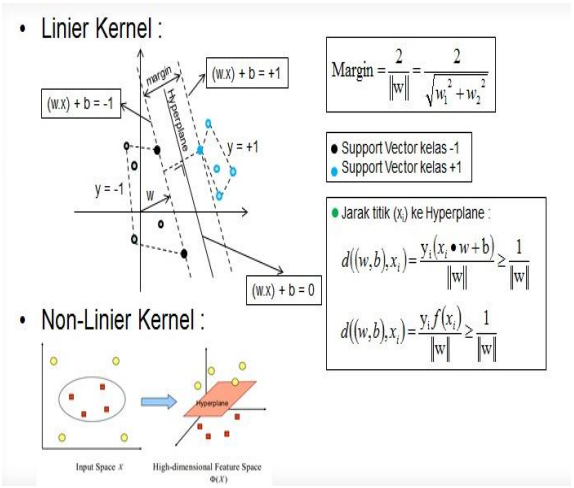
1. Tentukan jumlah klaster
2. Alokasikan data secara random
3. Hitung centroid / rata-rata dari data yang ada di masing-masing cluster
4. Alokasikan data ke centroid ke centroid terdekat.
5. Kembali ke langkah ketiga jika masih ada data yang berpindah cluster.

Euclidean distance space digunakan untuk menghitung jarak, hal ini karena hasil yang diperoleh merupakan jarak terpendek antara dua titik yang diperhitungkan. Adapun persamaannya sebagai berikut :

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left\{ x_{ik} - x_{jk} \right\}^2}$$

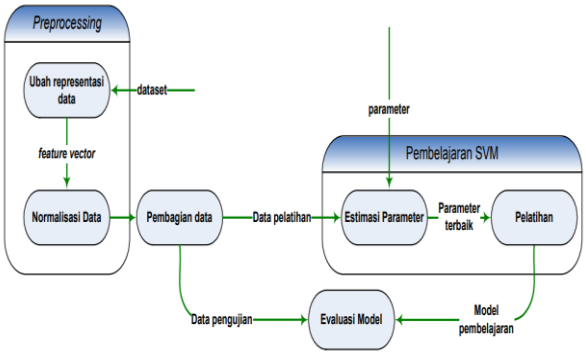
d*ij* = jarak objek i dan j
P = Dimensi data
Xik = Koordinat dari objek i pada dimensi K
Xjk = Koordinat dari objek j pada dimensi k

Support Vector Machine (SVM) adalah metode klasifikasi yang bekerja dengan cara mencari *hyperplane* dengan margin terbesar *Hyperplane* adalah garis batas pemisah data antar-kelas. Menurut J. Yunliang, et al [5] Margin adalah jarak antara *hyperplane* dengan data terdekat pada masing-masing kelas. Adapun data terdekat dengan *hyperplane* pada masing-masing kelas inilah yang disebut *support vector mechine*. Menurut J.Z.Liang [6] Pada dasarnya, SVM merupakan metode yang digunakan untuk klasifikasi dua kelas (*binary classification*). Pada perkembangannya, beberapa metode diusulkan agar SVM bisa digunakan untuk klasifikasi *multi-class* dengan cara mengombinasikan beberapa *binary classifier*.



Gambar 1. Visualisasi SVM

Unsupervised learning dengan SVM data pelatihan dan data pengujian adalah data yang sama. Selain itu, untuk proses pelatihannya dapat juga hanya menggunakan sebagian data dari data pengujian sehingga proses waktu pelatihan menjadi lebih singkat, tetapi hal ini mungkin menurunkan akurasi pada tahap pengujian.



Gambar 2. Pembelajaran dengan SVM.

Rapid miner adalah *tool* yang digunakan untuk penelitian *Text Mining* ini. RapidMiner sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan munggunakan bahasa java sehingga dapat bekerja di semua sistem operasi. RapidMiner sebelumnya bernama YALE (Yet Another Learning Environment), dimana versi awalnya mulai dikembangkan pada tahun 2001 oleh RalfKlinkenberg, Ingo Mierswa, dan Simon Fischer di Artificial Intelligence Unit dari University of Dortmund. RapidMiner didistribusikan di bawah lisensi AGPL (GNU Affero General Public License) versi 3. Hingga saat ini telah ribuan aplikasi yang

dikembangkan mengunakan RapidMiner di lebih dari 40 negara.

HASIL DAN PEMBAHASAN

Hasil penelitian ini menguji 150 judul karya ilmiah dosen dan mahasiswa STMIK Widya Cipta Dharma dengan tahun kegiatan 2014-2016. Dengan menerapakan metode SVM (Support Vector Mechine) serta optimasi dengan mengabungkan metode K-Means Clustering dan SVM. Dari 150 data tersebut dibuatlah data training sebannyak 70 % (105) data sebagai data training dan 30 % (45) data sebagai data training. Penelitian dilakuakan di STMIK Widya Cipta Dharma

Penelitian ini akan dilaksanakan melalui beberapa tahapan yaitu :

1. Menentukan dataset

Sebagai sumber data penelitian yaitu data tema karya ilmiah dosen dan mahasiswa tiga tahun dari tahun 2014-2016, sebanyak 150 Karyayang terdiri dari berbagai tema hasil karya ilmiah. Tool (Software) yang akan digunakan dalam penelitian ini adalah RapidManer dan sebagai pendukung pengolahan data mengunakan Microsoft Excel 2010.

2. Praprocessing Data

Tahap awal sebelum melakukan proses pengelompokan dokumen adalah mempersiapkan teks yang ada didalam dokumen. Pada tahap praproses ini dilakukan beberapa subproses agar dokumen dapat dipakai untuk melakukan proses pengelompokan. Subproses diantaranya yaitu:

a. *Tokenizer*, yakni proses yang bertujuan untuk memisah teks menjadi beberapa *token* berdasarkan pembatas berupa spasi atau tanda baca.

b. Proses selanjutnya adalah menghilangkan teks yang bersesuaian dengan teks yang terdapat pada daftar *stopword*, karena teks tersebut dianggap tidak dapat mewakili konten dokumen.

c. Kemudian pada teks yang masih tersisa dilakukan proses *stemming*, yaitu proses pengubahan teks menjadi bentuk dasarnya.

d. Selanjutnya, setiap kata tersebut disebut sebagai *term*. Nantinya setiap *term* akan didaftar dan diberi bobot.

e. Pembobotan masing-masing term dilakukan dengan metode TF-IDF (*Term Frequency – Inverse Document Frequency*). TF-IDF merupakan metode pembobotan *term* dengan menggunakan *termfrequency* (jumlah *term* yang terdapat pada tiap dokumen) serta *inverse document frequency* (*invers* jumlah dokumen yang memuat suatu *term*).

3. Pengelompokan Dokumen

Dari *k* model klasifikasi yang telah ada, maka dapat dilakukan klasifikasi dokumen baru. Pengujian dilakukan dengan mengelompokkan dokumen baru kedalam kelompok yang ada menggunakan tetangga terdekat dari *centroid* pada masing-masing kelompok. Setelah didapatkan kelompok yang sesuai maka dilakukan proses klasifikasi dokumen baru dengan model *SVM* pada kelompok yang bersangkutan.

4. Penentuan data Training dan data testing

Data *training* dan *testing* dalam penelitian ini diambil dari judul tugas akhir mahasiswa program studi S1 Sistem Informasi dan Teknik Informatika serta D3 Manajemen Informatika STMIK Widya Cipta Dharma, dimana setelah dijumlahkan akan di split menjadi 70% data *training* dan 30% data *testing*

5. Eksperimen dan Pengujian

Dari *k* model klasifikasi yang telah ada, maka dapat dilakukan klasifikasi dokumen baru. Pengujian dilakukan dengan mengelompokkan dokumen baru kedalam kelompok yang ada menggunakan tetangga terdekat dari *centroid* pada masing-masing kelompok. Setelah didapatkan kelompok yang sesuai maka dilakukan proses klasifikasi dokumen baru dengan model *SVM* pada kelompok yang bersangkutan

6. Evaluasi dan Validasi

Pada penelitian ini sebagai evaluasi dari model yang diusulkan, yaitu dengan menggunakan metode *cross validations* untuk mencari nilai akurasi yang kemudian hasil dari akurasi tersebut dievaluasi dengan cara membandingkan tingkat akurasi yang dihasilkan oleh model svm dengan menggunakan *k-means* dan dengan model svm tanpa *k-means*

Tabel 1. Daftar judul penelitian

No	Judul	Kategori
1	application library on sman 1 sebulu	Aplikasi & Sistem
…..	……..	…
150	implementation of home light control using arduino uno based	Hardware &jaringan

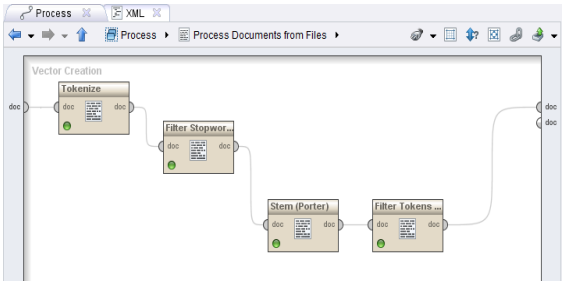
a. Pra Procesing data

sebelum data set diolah dan diujicoba dega mengunakan metode yang dipilih terlebih dahulu dataset dirubah kedalam bentuk file

bertipe .txt. Hal ini akan membantu memudahkan tools Rapid Miner dalam membaca dan mengolah data. Dataset yang sudah dirubah kedalam bentuk file .txt, dikelompokan penyimpanannya kedalam folder-folder yang penamaannya disesuaikan denga kategori masing-masing judul

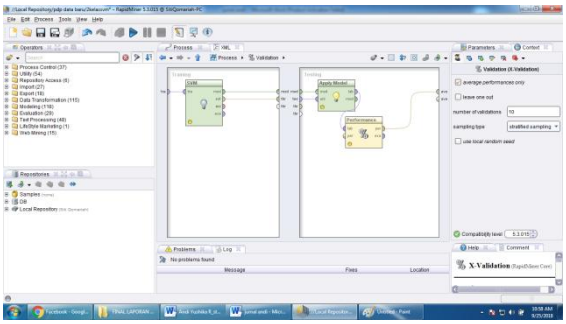
b. Procesing Data

Pada tahapan ini dataset yang sudah di lakukan tahapan pra prosesning akan di load pada tool rapi miner, loading data set. kemudian data yang sudah di loading dilakukan tahapan text procesing seperti token, steeming, filter. Seperti pada gambar 3 di bawah ini :



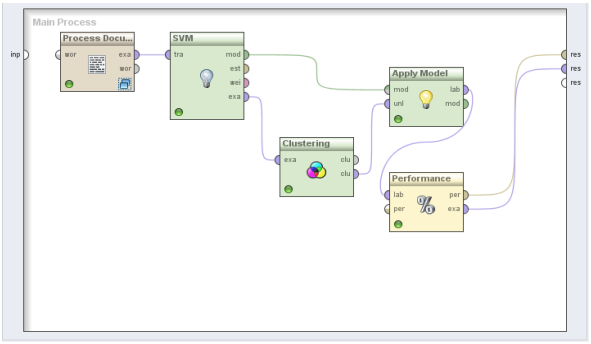
Gambar 3. Text processing

Penerapan metode dengan menggunakan *tool* rapid miner mengolah data dengan metode yang tersedia rapid miner, baik menggunakan metode SVM maupun K-Means Clustering dengan SVM. Pada gambar 4 terlihat pemilihan metode SVM kemudian dihubungkan dengan *Apply Mode dan Performance* untuk mengetahui tingkat akurasi dari metode tersebu.



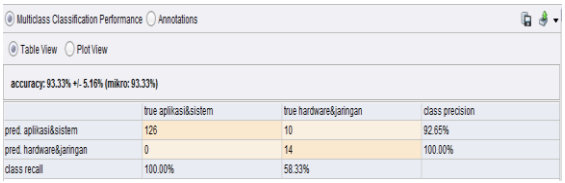
Gambar 4. Penerapan metode SVM

Pada gambar 5 pada *tool Rapid minner* data yang sudah di olah menjadi file dengan ekstensi *.txt kemudian dilakukan pengolahan data dengan *text Procesising* kemudian mengunakan metode SVM dan K-means Clustering selanjutnya *apply mode* dan *performance* untuk mengukur tingkat akurasi



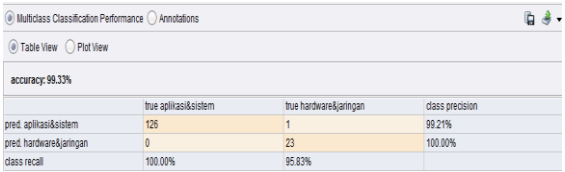
Gambar 5. Penerapan metode K-Means Clustering dengan SVM

Diantara mekanisme yang dapat dilakukan untuk mengukur validitas hasil klasifikasi adalah dengan menghitung nilai prediction dan recall. Perhitungan nilai precision akan mengukur tingkat kepastian (exactness) atau jumlah data testing yang diklasifikasikan dengan benar oleh model klasifikasi yang dibangun Perhitungan recall merupakan kebalikan dari precision. Recall mengukur sensitifitas atau rasio dari data untuk setiap label yang diklasifikasikan dengan benar terhadap data yang salah diklasifikasikan ke label lainnya (*misclassified*). Pada masing-masing hasil evaluasi dapat dilihat pada gambar berikut :



Gambar 6. Hasil Evaluasi metode SVM

Pada Gambar 6 hasil evaluasi menunjukan hasil akurasi model adalah 93.33% dengan class prediksi untuk hardware dan jaringan yang mencapai 100% tetapi class recallnya 58.33%. class prediksi untuk aplikasi dan sistem mempunyai nilai 92,65% dengan class recall sebesar 100%.



Gambar 7. Hasil Evaluasi K-Means Clustering dan SVM

Pada gambar 7 hasil evaluasi menunjukan hasil akurasinya sebesar 99,33% dengan rincian class prediksi untuk hardware dan jaringan sebesar 100% dan untuk aplikasi dan system sebesar 99,21%. Class recall untuk hardware dan jaringan sebesar 95,83% dan untuk apliksi dan jaringan sebesar 100%.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan hasil evaluasinya dapat digambarkan dalam bentuk grafik, hasil akurasi metode svm menghasilkan tingkatakurasi 93,33 % dan hasil akurasi k-means clustering dan svm sebesar 99,33%

SARAN

Penelitian ini berfokus hanya pada dua kelas diharapkan kedepannya bisa lebih banyak kelas sehingga jelas pengklasifikasian dari masing-masing tema karya tulis.

UCAPAN TERIMA KASIH

Terimakasih peneliti ucapkan atas hibah penelitian yang diberikan dari kemenristekDikti anggaran tahun 2018

DAFTAR PUSTAKA

[1] Gupta, N., "Text Mining for Information Retreival," Ph.D. thesis, Jaypee Institute of Information Technology

[2] Ian H Witten, Eibe Frank, and Mark A Hall, Data Mining PracticalMachine Learning Tools and Techniques. USA: Elsevier, 2011.

[3] Somantri,O. Wiyono,S. Dairoh., 2014, *Metode K-Means Untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Mengunakan Supprot Vector Machine (SVM)*. scientific journal of informatic vol 3 no 1 mei 2016 Hal 34-45

[4] Turban, E., Aronson, J.E., Liang, T.P., Introduction to Data Mining, Pearson, 2005.

[5] Yunliang, J., Qing, S., Jing, F., & Xiongtao, Z. 2010,. The Classification for E-government Document Based on SVM. In Web Information Systems and Mining (WISM), 2010 International Conference on (Vol. 2, pp. 257-260).

[6] J.Z. Liang. 2004 "SVM Multi-Classifier And Web Document Classification", Proceedings of the IEEE Third International Conference on Machine Learning and Cybernetics.