

# タグ付きコーパスを用いた動向情報とその要因の可視化

山本健一 殿井加代子 谷岡広樹

株式会社ジャストシステム

〒 771-0189 徳島市川内町ブレインズパーク

{kenichi\_yamamot,kayoko\_tonoi,hiroki\_tanioka}@justsystem.co.jp

## 概要

近年、電子化された情報の増加に伴い、ユーザの関心や興味に合致する情報に直接的かつ簡便にアクセスするための技術が求められている。このような要求に応える技術のひとつとして、我々は、動向情報の変化とその変化要因とを視覚的に表示するシステムを研究している。本稿では、動向情報として内閣支持率、要因として新聞記事を用いたシステムに関して報告する。本システムは、内閣支持率に関連する新聞記事を入力することにより、(1)内閣支持率の推移グラフを出力し、(2)ユーザの興味と見やすさを考慮し、内閣支持率の変化の大きい部分などにその変化の根拠となる要因をグラフ上に配置する、ことを特徴とする。

**Keywords:** 動向情報、情報可視化、TF・IDF、Support Vector Machine、Naive Bayes

## 1 はじめに

計算機の処理能力の向上や高速ネットワーク環境の普及に伴い、電子化された情報は増加の一途を辿っており、この傾向は今後も継続するものと思われる。そのため、ユーザの関心や興味に合致する情報に直接的かつ簡便にアクセスするための技術が求められている[5][10]。

このような要求に答える技術のひとつとして、我々は、動向情報の変化とその変化要因とを視覚的に表示するシステムを研究している。本稿では、動向情報として内閣支持率、要因として新聞記事を用いたシステムを開発することを課題とする。

本システムは、内閣支持率に関連する新聞記事を入力することにより、内閣支持率の変化とその変化の根拠となる新聞記事に関連付けるグラフを出力する。さらに、グラフ上には、ユーザの興味と見やすさを考慮し、内閣支持率の変化の大きい部分などに根拠となる要因を配置する。ここで、根拠となる新聞記事は、内閣支持率に関連する新聞記事とのコサイン距離に近い

記事を内閣支持率の変化の根拠とし、その重要語は、TF・IDFに基づくスコアを利用する。

このとき、ユーザの興味と見やすさについては、アンケート調査を行うことにより、要因を表示する位置や、その表示内容に関して有用な知見を得る。

以上のような構成により、内閣支持率の変化とその要因との関係を視覚的に表現し、かつユーザの関心や興味に合致する情報に効率的にアクセス可能なシステムが開発できたので報告する。

更に、Support Vector Machine や Naive Bayes を用いて動向情報の予測実験を行い、また動向情報の変化に寄与しやすい語の算出を行ったので、併せて報告する。

次節では、動向情報と要因について詳しく説明する。さらに第3節で、動向情報と要因の抽出手法に関して述べる。続いて第4節では、動向情報とその要因の表示方法に関するユーザ実験に関して報告する。そして第5節では、動向情報の予測実験に関して報告し、最後に第6節では、本稿のまとめと今後の課題に関して述べる。

## 2 動向情報とその要因

本節では、動向情報や要因とは何なのかを説明する。まず始めに動向情報に関して説明し、続いて要因に関して述べる。

### 2.1 動向情報とは

松下ら[5]は、動向情報とは、いくつかの統計量に関する時系列データを基にして、その変化を通時的に捉えて纏め上げるものであり、それは単に時系列データの羅列ではなく、ある観点の下で統合的に纏め上げることで得られるものである、としている。

我々はこの定義に従い、特に本稿においては1998年1月から1999年12月までの内閣支持率を研究対象の動向情報とする。

表 1: 検索結果（一部抜粋）

検索年月	検索結果タイトル
1998 年 1 月	新党支持率、5 党合計で「7 %」「離合集散」に期待薄く - - 毎日新聞社世論調査
	[特集] 世論調査「政治意識」(その 4 止) 質問と回答
	橋本龍太郎内閣 支持率、最低の 27 % 「景気」で不信心 - - 本社世論調査 【大阪】
2 月	[社説] 世論調査 橋本内閣の「不安」と「安心」
	[争点論点] 民友連結成 1 カ月 民主党代表・菅直人さん / 椎橋勝信 (その 2 止)
	いまの日本にふさわしい首相は... 菅直人氏、15 % でトップ - - 本社電話調査 【大阪】
3 月	自民 4 ポイント増 28 %、民主 3 ポイント増 8 % - - 政党支持率・毎日本社世論調査
	[特集ワイド] 渡部恒三・衆院副議長インタビュー 「さくらんぼの木」構想は !?
	[特集ワイド 1] クリントン・スキャンダルと米社会

## 2.2 要因とは

要因とは、動向情報に変化を与える「もの」とする。内閣支持率においては、様々な要因が考えられるが、それらはいずれも新聞記事に記載されているはずである。そこで我々は、内閣支持率に変化を与える新聞記事をその要因とすることとする。

## 3 動向情報とその要因の抽出

本節では、動向情報やその要因の抽出に関して述べる。まず始めに、動向情報の抽出に関して述べ、次に要因の抽出手法に関して説明する。

### 3.1 動向情報の抽出

我々は、MuST[1] オーガナイザから提供されたタグ付きコーパスを用いて内閣支持率の抽出を行うことを試みた。しかし、タグ付けの精度やその量が問題となり、1998 年 1 月から 1999 年 12 月までのすべての内閣支持率を抽出することはできなかった。

そこで、以下の 2 点の理由から人手により内閣支持率を与えることとした。

- インターネット上にはこれらの情報は纏めて保存されている [3] ことから、人手で与えるコストは十分に低いこと
- 今後タグ付けの精度やその量が十分になれば、タグ付けコーパスからの内閣支持率の抽出は容易に行えること

### 3.2 要因の抽出

我々は、内閣支持率の変化の要因となる新聞記事と、内閣支持率が記載された新聞記事とでは、使用される語の分布が近いと仮定した。そこで、MuST オーガナイザから配布された内閣支持率に関するタグ付きコー

パスをクエリとし、毎日新聞 2 年分のコーパスを検索した。検索結果の新聞記事が掲載された年月ごとに新聞記事を検索スコアの高い順にソートした。なお、検索に用いたシステムは、ベクトル空間法に基づいて我々が以前開発したシステム [11][12] である。結果の一部を表 1 に示す。このようにして検索された新聞記事を、その月の内閣支持率に変化を与えた要因とする。

次に、検索された新聞記事から重要語を抽出することにした。重要語の抽出は以下の手順により行われる。

ステップ 1: 同じ年月に掲載されたすべての新聞記事を 1 つのドキュメントと見なし、TF・IDF 値を算出する

ステップ 2: 算出された TF・IDF 値に基づいて、検索された新聞記事から重要語を算出する

このようにして算出された重要語を、グラフ上にタイトルを表示することが困難な場合などに使用する。ステップ 1 により算出した TF・IDF 値の一部を表 2 に示す。

## 4 ユーザ実験

抽出した動向情報や要因をどのように提示すれば良いのかを調査するためにアンケート調査を行った。まず始めに、要因の表示箇所と表示数に関する調査結果を述べ、次に、要因の表示内容に関する調査結果を述べる。

### 4.1 要因の表示箇所と表示数

初期状態で表示される要因の表示箇所に関して、次のようなアンケートを行い調査した。なお、アンケートは 61 人（男性：43 人、女性：18 人）のジャストシステム社員に対して行った。アンケート回答者の職種は、開発職、営業職、スタッフ職である。

表 2: TF・IDF 値算出結果（一部抜粋）

算出年月	ターム
1998 年 1 月	井坂理事
	民友連
	友会
	大統一会派
	紙本墨書
2 月	新井議員
	井坂前理事
	民友連
	パ社
3 月	リビンスキー
	歯止
	民友連
	商
	1 w
	キトラ

表 3: 要因表示箇所に関する調査結果

位置	#1 位 (割合)	#2 位以下 (割合)
A	1 (0.016)	4 (0.028)
B	0 (0.000)	2 (0.014)
C	2 (0.033)	20 (0.142)
D	2 (0.033)	13 (0.092)
E	0 (0.000)	0 (0.000)
F	4 (0.066)	15 (0.106)
G	45 (0.738)	15 (0.106)
H	5 (0.082)	50 (0.355)
I	0 (0.000)	1 (0.007)
J	1 (0.016)	2 (0.014)
K	0 (0.000)	7 (0.050)
L	1 (0.016)	11 (0.078)
計	61	140
総計		201
平均選択個数		201 / 61[人] = 3.30

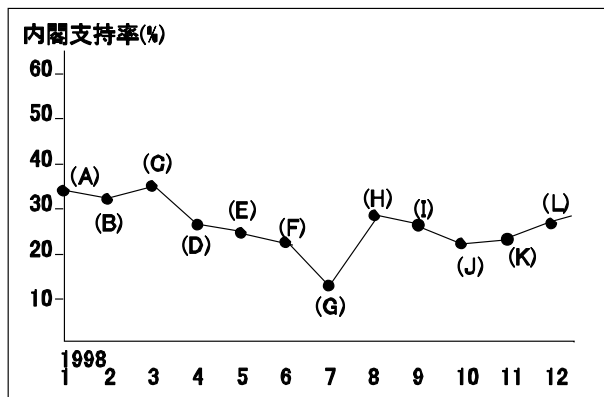


図 1: 内閣支持率のグラフ

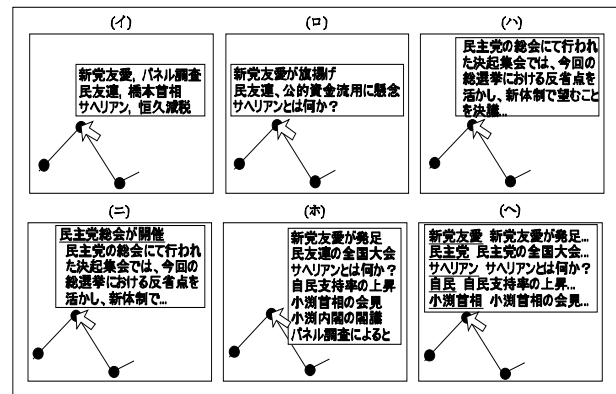


図 2: 要因の表示内容

質問 1: 図 1 は、内閣支持率の変動をグラフで表現したものです。このグラフ中の (A) から (L) でマークされた場所のうち、どの位置の詳細な情報が知りたいですか？(1) 見たい順に好きなだけ見たい位置の番号を記入してください。(2) その理由をなるべく具体的に記入してください。

アンケート結果を表 3 に示す。表 3、及びアンケート自由記述結果から以下のような傾向が分かった。

- 要因を知りたい箇所は、12 箇所中平均 3.3 箇所である
- 要因を知りたい箇所は、以下の 3 種類に分類できる
  1. 値の変化が大きい部分とその前後 (G、F、H)
  2. 値が最大の位置と最小の位置 (C、G)
  3. グラフの最初と最後 (A、L)

## 4.2 要因の表示内容

次に、要因の初期表示内容に関して、同様にアンケート調査を行った。アンケート回答者は、質問 1 と同じ 61 人のジャストシステム社員である。

質問 2: 図 2 における (イ) から (ヘ) は、マウスクリックによってある月の内閣支持率に関係がある情報を表示しています。(イ) から (ヘ) のうち、どの情報が見やすいですか？(1) 見やすいと思う順に 1 個以上記入してください。(2) その理由をなるべく具体的に記入してください。(どの表示形式からも、さらにマウスクリックで詳細表示が可能です。)

アンケート結果を表 4 に示す。表 4、及びアンケート自由記述結果から以下のような傾向が分かった。

表 4: 要因表示内容に関する調査結果

内容	#1 位 (割合)	#2 位以下 (割合)
イ	4 (0.066)	17 (0.205)
ロ	20 (0.328)	15 (0.181)
ハ	1 (0.016)	5 (0.060)
ニ	16 (0.262)	15 (0.181)
ホ	4 (0.066)	18 (0.217)
ヘ	16 (0.262)	13 (0.157)
計	61	83
総計	144	
平均選択個数	144 / 61[人] = 2.36	

- 1つのポップアップウィンドウ内には、3行くらいがよい（ロとホの比較より）
- 内容を簡潔に表すタイトルやキーワードがある方が分かりやすい（ハとニ、ヘとホの比較より）
- イのようなキーワードの羅列は分かりにくい（ただし、分析に馴れている人はキーワードの羅列でも良いとの評価も多かった）
- 内容をただ表示するのは、分かりにくい（ハの結果より）
- “...”で終わると、クリックできることがよく分かる

## 5 動向情報の予測

ここでは、Support Vector Machine と Naive Bayes を用いた内閣支持率の予測実験に関して述べる。また、内閣支持率の変化に寄与しやすい語に関しても併せて報告する。

### 5.1 実験方法

内閣支持率の変動は、前の月の新聞記事の内容に依存すると仮定し、内閣支持率の予測問題を以下の様に定式化した。

$$y_i = \phi(x_{i-1}) \quad (1)$$

ただし、 $i$  は年月を表す添え字とし、 $x_{i-1}$  は、 $i-1$  を表す年月に発行された新聞記事から抽出された  $n$  次元の特徴ベクトルとする。また、 $\phi$  は入力された  $n$  次元の特徴ベクトルを 1 または 0 に写像する関数とする。更に、

$$y_i = \begin{cases} 1 \dots \text{内閣支持率の上昇} \\ 0 \dots \text{内閣支持率の下降} \end{cases} \quad (2)$$

とする。

ここで本実験においては、 $x_{i-1}$  は、 $i-1$  を表す年月に発行された新聞記事から抽出された名詞句を次元として、TF または TF · IDF の値をその次元の値とした。更に、 $i-1$  を表す年月に発行された新聞記事から、MuST オーガナイザから配布された内閣支持率に関するタグ付きコーパスをクエリとして検索し、検索結果の上位のみを利用した場合も比較実験した。なお、検索に用いたシステムは、ベクトル空間法に基づいて我々が以前開発したシステム [11][12] である。

次に関数  $\phi$  に、Support Vector Machine を用いた場合と、Naive Bayes を用いた場合についてそれぞれ説明する。

#### 5.1.1 Support Vector Machine

SVM(Support Vector Machine) は、マージン最大化の戦略に基づく分類器であり、その高い汎化能力から広く自然言語処理の分野で用いられている [20, 19]。

我々は、LIBSVM[13] を SVMs のライブラリとして使用し実験を行った。実験パラメータとして、カーネル関数に RBF カーネルと Linear カーネルを利用し、C 値は 10 とした。

#### 5.1.2 Naive Bayes

NB(Naive Bayes) は、ベイズの定理に基づく分類器 [14, 15, 16, 17, 18, 19] であり、ここでは、 $p(y_i|x_{i-1})$  を最大にする  $y_i$  を探せばよい。

$$p(y_i|x_{i-1}) \propto p(y_i) \prod_{j=1}^n p(w_j|y_i) \quad (3)$$

ただし、ここで  $w_j$  は  $x_{i-1}$  の  $j$  次元成分のベクトルとする。

### 5.2 結果

表 5 に 1998 年 1 月から 1999 年 12 月までの 24 ヶ月分の毎日新聞コーパスを用いて、leave one out で評価した結果を示す。表 5 において、ALL は、その期間のすべての新聞記事を対象とした場合の結果であり、Search-ALL は、MuST オーガナイザから配布された内閣支持率に関するタグ付きコーパスをクエリとして検索し、全検索結果を対象とした場合の結果である。また、Search-[数値] は、検索結果の上位 [数値] 件を対象とした場合の結果を示している。

表 5: 内閣支持率の予測実験結果

	NB TF	SVM TF RBF	SVM TF Linear	NB TFIDF	SVM TFIDF RBF	SVM TFIDF Linear
ALL	45.8	0	50.0	50.0	0	50.0
Search-ALL	54.2	0	50.0	50.0	8.3	54.2
Search-5000	54.2	0	58.3	54.2	8.3	54.2
Search-1000	54.2	0	58.3	45.8	12.5	45.8
Search-800	58.3	0	58.3	50.0	12.5	45.8
Search-500	62.5	0	58.3	50.0	12.5	50.0
Search-300	62.5	12.5	58.3	54.2	16.7	62.5
Search-100	70.8	45.8	66.7	62.5	25.0	50.0
Search-50	66.7	58.3	62.5	62.5	25.0	50.0
Search-30	66.7	62.5	66.7	54.2	25.0	50.0
Search-10	70.8	58.3	54.2	66.7	25.0	50.0
Search-5	70.8	79.2	70.8	58.3	50.0	54.2
Search-1	70.8	45.8	54.2	75.0	62.5	62.5

表 5 より、学習データは検索により絞り込んだ方が  
良い結果を得られることが分かった。また、今回の実  
験においては、各次元の値は TF・IDF よりも TF が安  
定している傾向にあると言える。

### 5.3 分析

ここでは、MuST オーガナイザから配布された内閣  
支持率に関するタグ付きコーパスをクエリとして検索  
し、検索結果上位 100 件を用いて、Naive Bayes で TF  
を用いて学習した場合の単語  $t$  が内閣支持率に与える  
影響  $cause(t)$  を以下の様に計算した。

$$cause(t) = \log \left( \frac{p(t|y=1)}{p(t|y=0)} \right) \quad (4)$$

従って、 $cause(t)$  が正の方向に大きいほど内閣支持  
率を上昇させる傾向にあり、逆に  $cause(t)$  が負の方向  
に大きいほど内閣支持率を下降させる傾向にある。

表 6 に内閣支持率上昇、下降に寄与する単語の上位  
を示す。表 6 より、上昇に寄与する単語はある程度上  
手く抽出できているが、下降に寄与する単語は数詞な  
どノイズが混ざっていることが分かる。

## 6 おわりに

本稿では、動向情報に関する要因の抽出手法と、動  
向情報とその要因の表示方法に関するアンケート調査  
結果に関して報告した。

要因の抽出では、動向情報が記載された新聞記事と  
要因が記載された新聞記事とでは、使用される単語の  
分布が近いという仮定の下に、ベクトル空間上で動向

表 6: 内閣支持率の変化に寄与する語

上昇に寄与	スコア	下降に寄与	スコア
号外	6.585	医	-6.857
再選挙	6.557	ラナリット	-6.343
トキ	6.400	農	-6.308
道府県議選	6.312	6 法案	-6.271
告示前	6.274	共新	-6.234
統一選	6.129	前期	-6.193
三上満	6.060	セン	-6.174
悟	6.060	6 0 後期	-6.044
舛添	6.036	横井	-5.996
4 グループ	6.012	検診	-5.866
菊田医師	5.934	3 . 6	-5.809
国連軍	5.934	3 . 9	-5.748
j 3 0	5.849	ポルポト派	-5.716
研	5.788	4 . 3	-5.716
玉井	5.724	自現	-5.684
告示後	5.724	大統一会派	-5.684
都知事	5.690	3 . 7	-5.650
1 5 区	5.655	4 . 2	-5.650
健治	5.619	4 . 4	-5.615
真須美	5.619	自新	-5.615

情報が記載された新聞記事とコサイン距離が小さい新  
聞記事を要因とした。また、要因とされた新聞記事か  
らのキーワード抽出は、1 月分の新聞記事を 1 ドキュ  
メントとみなした IF・IDF 値に基づき行った。

動向情報とその要因の表示方法に関するアンケート  
調査からは、ユーザが動向情報に関する要因を知りた  
いのは、動向情報を示したグラフ中の変化が大きい部  
分とその前後、最大位置と最小位置、及び最初と最後  
の 3 つに分類できることが分かった。また、要因の表  
示方法としては、1 ウィンドウに 3 つぐらいのトピッ

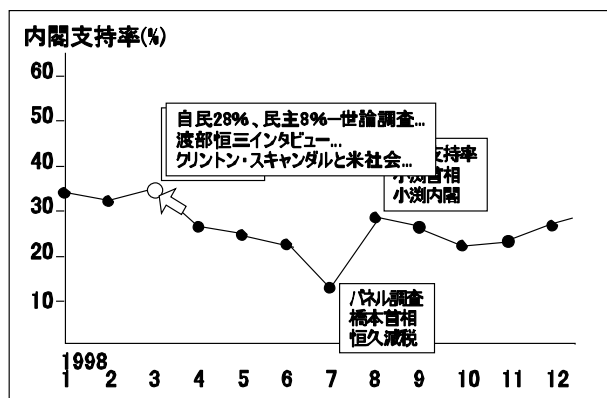


図 3: システムの出力例

クを簡潔に示したものや、要因をタイトルやキーワードと共に表示すれば良いことが分かった。

動向情報の予測実験では、Naive Bayes を用いて 70% 程度、Support Vector Machine を用いて 80% 程度の精度で正解することができた。また、内閣支持率の変化に寄与する語の収集も自動的に行うことができた。

本稿で述べたシステムの出力例を図 3 に示す。図 3 は、初期状態で表示されている内閣支持率とその要因から、ユーザが気になる箇所をクリックし、さらに詳細な情報を選択しているところを示している。

今後の課題として、以下のようなことが挙げられる。

- 要因の正当性の検証
- 動向情報の抽出実験
- 複数の動向情報を重ね合わせたグラフ（内閣支持率と株価、円相場など）
- 動向情報の予測実験に関する追加実験

## 謝辞

タグ付けコーパス、毎日新聞コーパスを提供して下さった MuST オルガナイザ、ラウンドテーブルミーティングやメーリングリストにおいて有益なディスカッションをして下さった方々、動向情報や要因の表示方法に関するアンケートに快く協力して下さったジャストシステム社員の方々に感謝します。

## 参考文献

- [1] 加藤恒昭, 松下光範, 神門典子, 動向情報の要約と可視化に関するワークショップ ホームページ, <http://must.c.u-tokyo.ac.jp/>
- [2] 加藤恒昭, 松下光範, 平尾努, 動向情報の要約と可視化に関するワークショップの提案, 情報処理学会自然言語処理研究会, 2004-NL-164 (15), pp.89-94, 2004.
- [3] 株式会社テレビ朝日 ニュースステーション, 内閣支持率, <http://www.tv-asahi.co.jp/n-station/research/naika-ku.html>
- [4] 加藤恒昭, 松下光範, 平尾努, 神門典子, 評価なきワークショップの試み — 「MuST: 動向情報の要約と可視化に関するワークショップ」を例に —, 言語処理学会全国大会併設ワークショップ「評価型ワークショップを考える」, 2005.
- [5] 松下光範, 加藤恒昭, 動向情報に基づく情報可視化の基礎検討, 人工知能学会第 19 回全国大会, 2005.
- [6] 難波英嗣, 国政美伸, 福島志穂, 相沢輝昭, 奥村学, 文書横断文間関係を考慮した動向情報の抽出と可視化情報処理学会自然言語処理研究会, NL-168, pp.67-74, 2005.
- [7] 難波英嗣, WWW 上のテキスト情報の知的統合, 『人工知能学会誌』, 19 巻 3 号, 2004.
- [8] 難波英嗣, 複数テキスト情報の可視化: 研究事例の紹介, 電子情報通信学会 Web インテリジェンスとインタラクション研究会, WI2-2005-28 ~ 49, pp.109-115, 2005.
- [9] 奥村学, 難波英嗣, 知の科学 テキスト自動要約, オーム社, 2005.
- [10] 福本淳一, 天野真家 (編), 特集 自然言語による情報アクセス技術, 情報処理, Vol. 45, No. 6, pp.561-585, 2004.
- [11] Hiroki Tanioka and Kenichi Yamamoto, A Distributed Retrieval System for NTCIR-5 Patent Retrieval Task, *The 5th NTCIR Workshop Meeting*, 2005.
- [12] Hiroki Tanioka, Kenichi Yamamoto and Takashi Nakagawa, A Distributed Retrieval System for NTCIR-5 WEB Task, *The 5th NTCIR Workshop Meeting*, 2005.
- [13] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines., 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [14] P. Domingos and M. J. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103-130, 1997.
- [15] D. Hand and K. Yu. Idiot's bayes - not so stupid after all? *International Statistical Review*, 69:385-399, 2001.
- [16] R. Irina. An empirical study of the naive bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 2001.
- [17] M. E. Maron. Automatic indexing: An experimental inquiry. *J. ACM*, 8(3):404-417, 1961.
- [18] M. Mozina, J. Demsar, M. W. Kattan, and B. Zupan. Nomograms for visualization of naive bayesian classifier. In *PKDD*, pages 337-348, 2004.
- [19] D. G. S. Richard O. Duda, Peter E. Hart. *Pattern Classification Second Edition*, chapter 2.12,5.11,9.6.2. John Wiley & Sons Inc, 2000.
- [20] V.N.Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.