第1回教員ゼミ

A Survey of Emerging Trend Detection in Textual Data Mining

From “Survey of Text Mining”

佐々木　稔

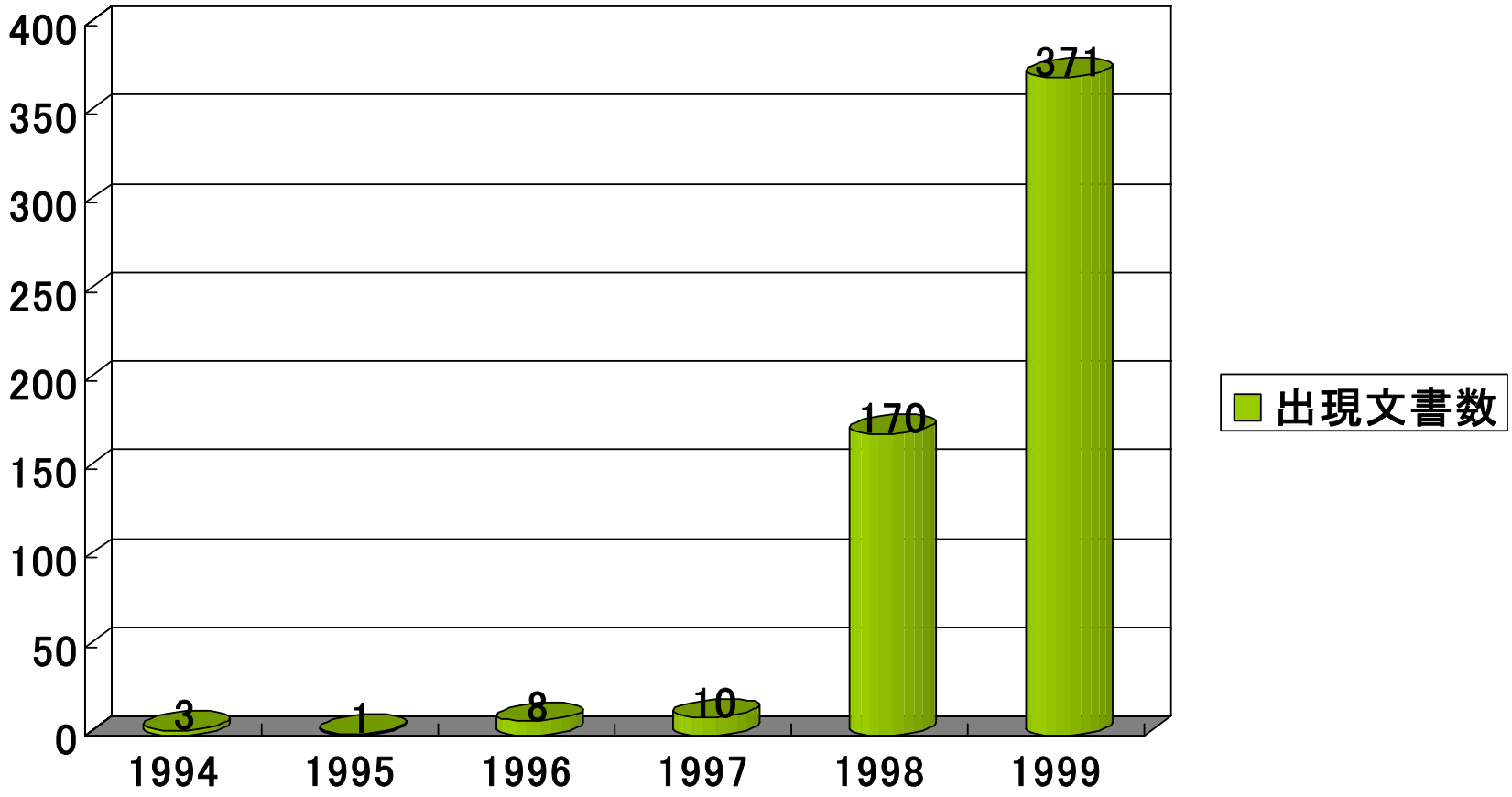
目次

- はじめに
- ETDシステムの研究事例
- 商用ETDシステム

はじめに(1/2)

- テキストデータからのトレンド(動向)の検出
- トレンド(動向)とは？
  - 時間と共に関心や有益さに変化のあるトピック
    - XML、Ajax、ライブドア、ワールドカップなど
- 動向分野を必要とする対象
  - 特定分野の動向を監視する個人、企業
    - 技術、関連文献のレビューなど

例:“XML”についての動向分析



はじめに(2/2)

- トレンド検出システム
  - 全自動:コーパス入力でトピックリストを出力
  - 半自動:トピックは手入力、出現傾向を出力
- システムに必要なもの(今回のポイント)
  - 言語学、統計学上の特徴
  - 学習アルゴリズム
  - 学習データ、テストデータの作成
  - 評価方法
- 動向の評価には専門家知識が必要

ETDシステムの評価データ

- Topic Detection and Tracking(TDT) Project
  - ニュース記事とイベント内容の組
  - 各組には判定結果が付与
  - 3種類のコーパスが存在(TDT-Pilot, TDT2, TDT3)
- INSPEC データベース
  - 工学論文の概要
  - アメリカ合衆国の特許

評価データの例

Story Description	Event	Relevance Judgment
Story describes survivor's reaction after Oklahoma City Bombing	Oklahoma City Bombing	Yes
Story describes survivor's reaction after Oklahoma bombing	US Terrorism Response	No
Story describes FBI's increases use of surveillance in government buildings as a result of the Oklahoma City Bombing	Oklahoma City Bombing	Yes
Story describes FBI's increased use of surveillance in government buildings as a result of the Oklahoma City Bombing	US Terrorism Response	Yes

Technology Opportunities Analysis (TOA) (1/2)

- 入力データと属性（緑文字は手入力）
  - キーワードリスト（*n*-gram）
  - 頻度、共起頻度、日付
  - 分野、分野内でのキーワード頻度
- 手順
  1. 特定分野から文書抽出
     - 専門家によるキーワードリストが必要
     - キーワードを組合せて質問を作り、適切な文書を検索
  2. Bibliometrics情報の解析
     - 単語頻度、日付、単語共起、引用、出版情報

Technology Opportunities Analysis (TOA) (2/2)

- 出力形態
  - 頻度統計（表、ヒストグラム、比率）
  - グラフ（両対数グラフ、Fisher-Pry曲線）
  - 技術マップ
- 評価
  - 表示結果を確認し、ユーザが評価
  - ユーザビリティ、注目クラスタの表示が有用

CIMEL: Constructive, Collaborative Inquiry-Based Multimedia E-Learning

- 協調学習のためのマルチメディアフレームワーク
  - 専門家ではなく、個人利用を対象
- 動向検出
  - Web資源を利用
  - 学生に向けてマルチメディアガイドを作る
- 特徴
  - 初期検出された動向に結果が集中する

CIMEL: Constructive, Collaborative Inquiry-Based Multimedia E-Learning

- 入力データ
  - Web資源
- 属性（緑文字は手入力）
  - 主題、動向候補
  - 主題を立証する事項
  - 検索項目（ ”XML ＜and＞ novel” など）
  - 日付、主題の頻度（出現ページ数）
  - 立証する事項の頻度
  - 立証する事項を含む行、または段落

CIMEL: Constructive, Collaborative Inquiry-Based Multimedia E-Learning

- 処理内容
  1. 調査分野を設定
  2. 最近の会議、ワークショップを設定
  3. 内容を調べて、動向候補リストを作成
  4. Web検索などで動向候補を評価
  5. 評価できない候補は、INSPECデータベースで検証
- 検証作業での利用属性
  - 年ごとの著者頻度、論文数、
  - 連名著者の組合せ数、開催地の使用頻度

CIMEL: Constructive, Collaborative Inquiry-Based Multimedia E-Learning

- 評価
  - 動向検出課題
    - オブジェクト指向プログラミングの「継承」
  - グループ1：CIMELによるマルチメディアチュートリアル
  - グループ2：ヒントなしで検索
  - 精度（正解率）で評価
    - グループ1の精度がグループ2を大幅に上回った

TimeMines

- 特徴
  - 文書データへの日付タグの付与
  - 時間軸上での重要なトピック表示が可能
  - ある時間間隔における適切なトピックの抽出
  - 最も重要な情報のみ表示
  - 時間に不変な出現率を持つ属性の分布からモデル作成

TimeMines

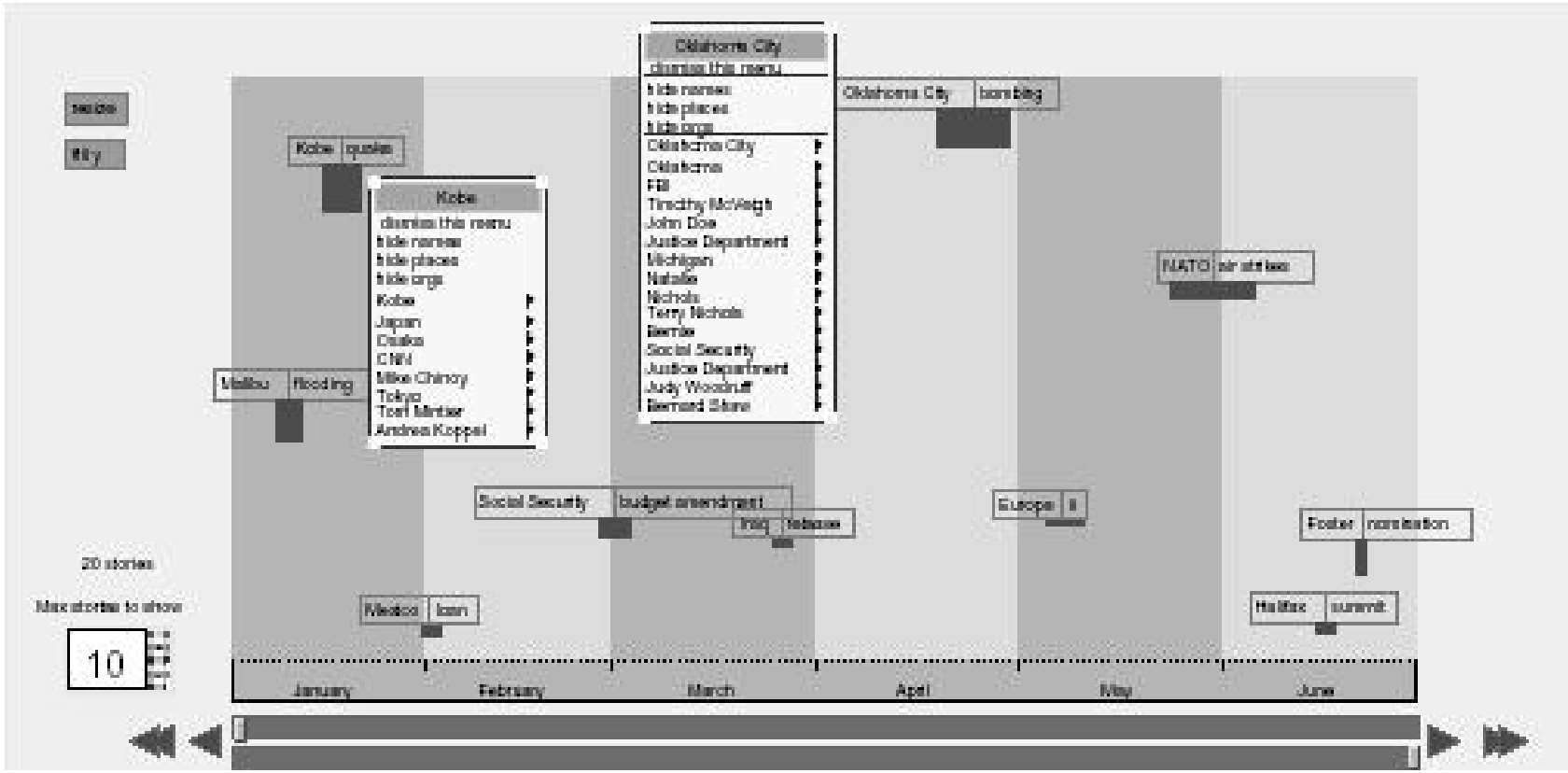
- 入力データ
  - TDT、TDT-2コーパス（日付タグ、品詞タグ付き）
- 属性
  - 固有表現（人名、場所、組織）
  - 「（名詞|形容詞）*名詞」となるパターン
  - 出現の有無：
    - 固有表現、n-gramが文書内にあれば “True”
    - なければ “False”
  - 日付

TimeMines

- 処理内容
  - 最も重要な特徴を抽出
    - 仮説検定に基づく統計モデル
    - すべての特徴が時間により分布が変化しないと仮定
    - モデルに合う特徴は捨てる
    - モデルに合わない特徴を候補として残す
    - 一定期間内で同じ分布を示す特徴をまとめる
  - 期間内で多くの特徴を持った少数のトピックを抽出

TimeMines

出力画面



New Event Detection

- 初出の話題を検出
  - ニュース記事から単語を抽出、過去の記事から順に比較
  - 比較
    - 単一パスクラスタリング
  - 内容（単語集合）をデータベース化
  - 新しい記事の処理でデータベースを検索
    - マッチすれば、過去に関連記事あり
    - マッチしなければ、新しい記事

New Event Detection

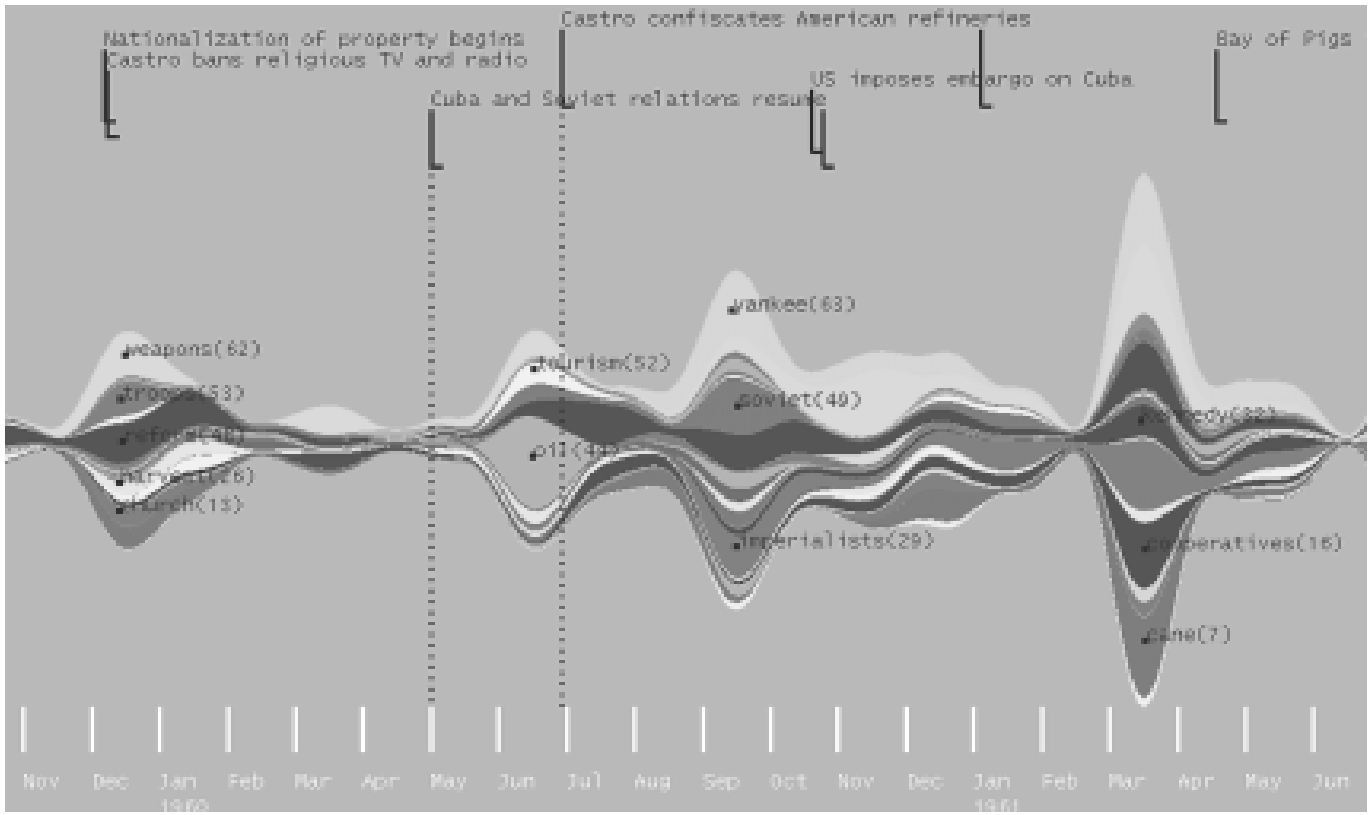
- 入力データ
  - TDTコーパス
- 属性
  - 単語ユニグラム
  - 記事内での単語頻度
  - 記事中の全単語数
  - 記事中の平均単語数
  - 単語の出現する記事数
  - 話題数
  - 日付

New Event Detection

- 処理内容
  - 単一パスクラスタリング（Single-pass method）
    1. いくつかの記事を初期クラスタとする
    2. 新しい記事とクラスタとの類似度を計算
    3. 最も高い類似度のクラスタに対して、
       - 閾値以上の類似度ならば、そのクラスタに割り当てる
       - 閾値以下ならば、新しいクラスタに割り当てる
    4. すべての要素が割り当てられるまで2～3を繰返す

ThemeRiver

- コーパス内の主要な話題を要約
- 時間ごとの話題の変化を「川」で表現



ThemeRiver

- 入力データ
  - 40年間のFidel Castroの演説、記事などのデータ
- 属性（緑文字は手入力）
  - 単語ユニグラム
  - 時間間隔における単語の出現する文書数
  - 日付

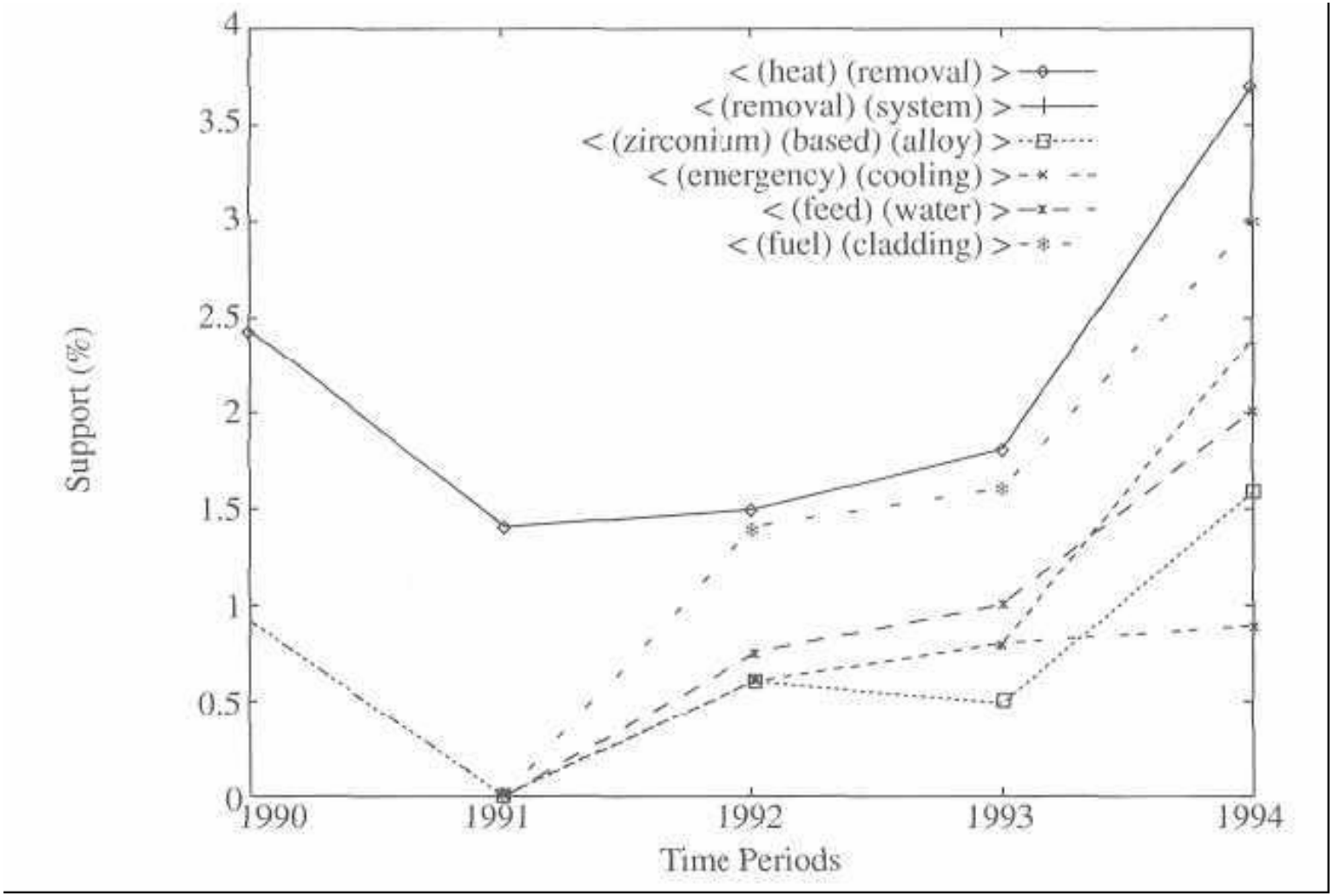
PatentMiner

- 特許データからの傾向発見システム
- データ
  - アメリカ合衆国特許
- 処理
  - フレーズ抽出（Sequential Pattern Mining）
    - 頻繁に共起する単語をまとめて、ひとつの話題として扱う
  - 傾向検出（Shape query）
    - 頻度の時間変化などの分布形状から話題を検出する

PatentMiner

- 属性（緑文字は手入力）
  - フレーズ（単語*n*-gram）
  - 最小間隔（単語、文、段落、章）
  - 最大間隔（単語、文、段落、章）
  - 時間窓（フレーズの単語数）
  - フレーズの出現確率
  - 日付
  - 時間における話題の変化

PatentMiner



Hierarchical Distributed Dynamic Indexing (HDDI)

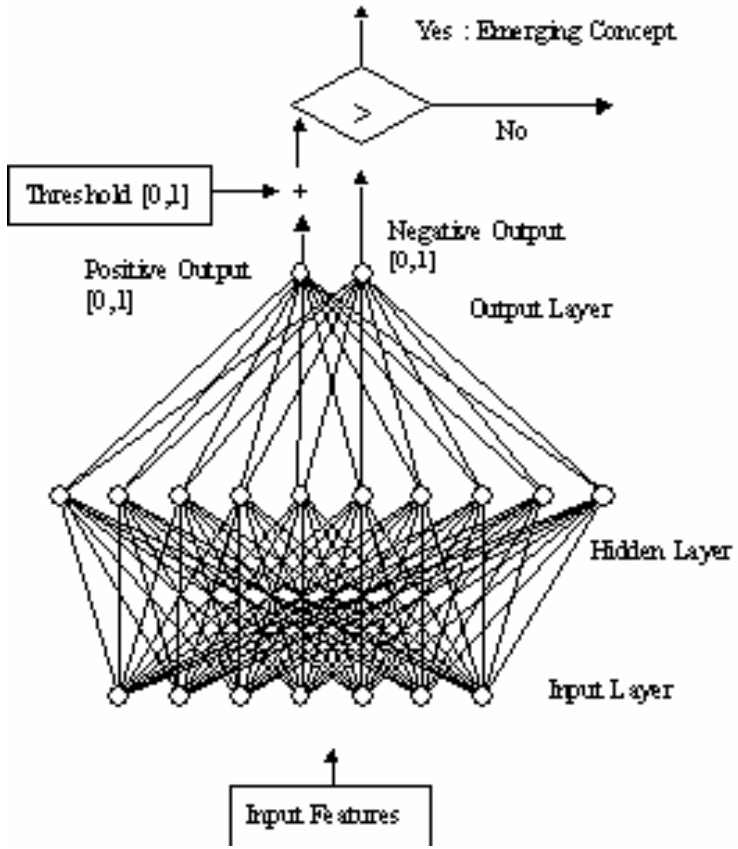
- 意味的に類似したクラスタの構築から話題を検出
- 非線形ニューラルネットワーク
  - クラスタサイズ、属性の時間変化を学習
- 学習データ
  - プロセッサ、パイプライン処理の出願特許3年分
  - 年ごとに内容により分類
- C4.5の利用により、同精度の処理が高速で検出

Hierarchical Distributed Dynamic Indexing (HDDI)

- 属性
  - フレーズ（正規表現により抽出）
  - 全文書中におけるフレーズの出現数
  - フレーズペアの共起頻度
  - 類似度
  - 平均値
  - 標準偏差

Hierarchical Distributed Dynamic Indexing (HDDI)

- 学習方法
  - 学習データを2種類に分類
    - 一定期間内のデータ
    - フレーズ間関係の統計値
  - ニューラルネットワーク
    - 逆誤差伝播法
      （back-propagation）



商用ETDシステム

- ETDに関心のある企業向けシステム
  - 内容の提供
  - 多目的な情報検索
  - クラスタリング

1. Autonomy
   - データを内容によりクラスタリング
   - クラスタの傾向から話題抽出
     - 現在と過去のクラスタを比較

商用ETDシステム

2. SPSS LexiQuest
   - キー概念と概念間の関係を抽出
   - キー概念抽出
     - 辞書ベースの言語解析
     - 統計的な類似度
   - 概念間の関係をマップ上に配置

商用ETDシステム

3. ClearForest
   - 大規模データから関連情報とその要約を抽出
     - 時間軸上での話題内容の流れ
     - 企業、人、イベントの関係を視覚的に表示
   - リアルタイムに更新が可能
     - 新製品情報、管理の変更、新技術など
     - 情報収集はWWW検索エンジンを利用
   - 「実体」と「事実」の抽出
     - 「事実」は「実体」間の関連
     - 機械学習により抽出