SUMMARIZING DOCUMENTS USING FRACTAL TECHNIQUES

María Dolores Ruiz & Antonio B. Bailón

{mdruiz,bailon}@decsai.ugr.es

UNIVERSITY OF GRANADA (SPAIN)
DEPT. COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE
Motivation

- A lot of information is accessible
- Most of this information is structured
- Necessity of automatic summarization in a fast and precise way

Solution: New models of automatic summarization using the structure of documents
Techniques used in automatic summarization

There are mainly three techniques:

▪ sentence extraction,
▪ bag-of-words headline generation
▪ document compression

We are going to use sentence extraction.
Sentence extraction

- Starts with the works made by Luhn and Edmundson
- The document is considered as a sequence of sentences
- They rank the sentences according to some salient features:
  - Thematic feature
  - Heading feature
  - Location feature
  - Cue feature
Our Proposal

- Taking into account the hierarquical structure of documents.
- Adapting the features to the structure.
- Improving the model proposed by Yang and Wang: *Fractal Summarization*.
- Removing its shortcomings proposing a different propagation formula.
- Using the fractal dimension of text documents to achieve a better performance.
Structure in Summarization Techniques (I)

- The document is considered as a tree taking into account its structure:
Structure in Summarization Techniques (II)

- **Features Adaptation to the structure:**
  - *Thematic feature:* is based in the frequency of words depending of the level (*tfidf* adaptation).
  - *Heading feature:* terms in higher levels are more important.
  - *Location feature:* assigns different values depending on the position of a node.
  - *Cue feature:* gives more importance to bonus words in a dictionary (conclusion, important...)

- The total weight is a normalized weighted sum of the above features.
Fractal Summarization (I)

- It was first proposed by Yang and Wang in 2003.
- It is based on the idea of Fractal View for controlling information display of H. Koike (1995).
- We have modified it to achieve a best perform of the method.
- We have introduced the *Fractal Dimension of a text document* in the method getting better results in summarizing web pages.
Fractal Summarization (II)

- The idea of Fractal View for controlling information display is based on a propagation formula.

\[
\begin{align*}
F_{v_{\text{root}}} &= 1 \\
F_{v_x} &= F_{v_{p(x)}} \times CN_{p(x)}^{-1/D}
\end{align*}
\]

C = 1, D = 1
Our proposal is taking the propagation formula:

\[
F_{V_{\text{root}}} = 1
\]

\[
F_{V_x} = F_{V_{\rho(x)}} \times \left( \frac{w(x)}{\sum_{y \in \rho(x)} w(y)} \right)^{\frac{1}{D(x)}}
\]

Where \( D(x) \) is the fractal dimension of the node \( x \) with all of its child nodes. (Considering the node \( x \) as the root)
What is a Fractal?

“fractal is a shape made of parts similar to the whole”

Mandelbrot

This definition uses the concept of Self-similarity.

A tree is an example of fractal:
Fractal Dimension

- Everybody knows the Euclidean dimension: a point has dimension 0, a line has dimension 1 and a plane has dimension 2.
- But some objects don't have an integer dimension, for example the previous tree has dimension:

\[
D = -\log_{6/10} 2 \approx 1.36
\]

- \(2\) = number of branches
- \(6/10\) = Scale factor (every branch is \(6/10\) smaller than the previous branch)
- This is a general formula for this case of fractal. But sometimes it is not possible to calculate the exact fractal dimension.
Methods for calculating fractal dimension

- Compass Method (only for curves)
- Box-Counting Method (only for low-dimensional sets)
- Other methods based in particular objects
- Grassberger and Procaccia method

We use the last method to calculate the fractal dimension of text documents.
Fractal dimension of text documents (I)

- We transform the document into a tree basing in its taxonomy
- We call node to every part of the document.
- We compute how similar are two nodes using a distance function based in the cosine of two vectors.

\[
dist(x, y) = 1 - \cos(x, y) = 1 - \frac{\sum_{t_i \in x} w_{ix} \sum_{t_i \in y} w_{iy}}{\sqrt{\sum_{t_i \in x} w_{ix}^2} \sqrt{\sum_{t_i \in y} w_{iy}^2}}
\]
Fractal dimension of text documents (II)

- We call:
  - \( N \) = number of nodes,
  - \( R \) = max. number of levels (words are the highest level)

- Suppose \( x_i = x_1, \ldots, x_n = x_j \) is the shortest path between the nodes \( x_i \) and \( x_j \), we define
  \[
  \| x_j - x_i \| = \sum_{l=1}^{n-1} \frac{\text{dist}(x_l, x_{l+1})}{2R}.
  \]

- Then, we calculate for each \( r_k = \frac{(2R-k)}{(2R)} \) with \( k = 1, \ldots, 2R-1 \) using the Grasberger and Procaccia method:
  \[
  C_m(r_k) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} I(\| x_j - x_i \| \leq r_k)
  \]

- And the last step is computing the slope of the regression line that points \( (\ln(C_m(r_k)), \ln(r_k)) \) form.
- This slope is the fractal dimension: \( D \).
Fractal Summarization Algorithm

Input:

- **Compression ratio.** \( r = \frac{\text{sentences in the summary}}{\text{sentences in the document}} \) (usually 0.2)
- **Threshold value** = max. number of sentences to extract from the same node. (3-5)
- **Quota** = sentences of the summary.

Steps:

- Transform the document into a tree.
- Compute the fractal dimension of each node of the document.
- **Calculate for each node:** \( \text{quota}_x = F_{v_p(x)} \times \text{quota}_{p(x)} \)
- **Fractal value:** we propose a new formula taking into account the fractal dimension of the document.
Propagation in Fractal Summarization

Document (Title)

Chapter 1
Quota=6

Chapter 2
Quota=10

Chapter 3
Quota=24

Section 1
Quota=3

Section 2
Quota=3

Subsection 2.1

Paragraph

Paragraph

Paragraph

Sentence

Sentence

Sentence

Sentence
Experiments and Results (I)

- We have done experiments with web pages with a good structure.
- We use the tags <H1>, <H2>,…, and the markup tags </H1>, </H2>, for organizing their structure.
- We build a tree using the mentioned structure.
- We consider at the last level only the terms with significance, dropping all the articles, ordinal and cardinal numbers, the verb to be, prepositions, pronouns, conjunctions, etc.
- We take the keywords of the document as the cue dictionary.
Experiments and Results (II)

- The method achieves a good performance with pages that have several levels of granularity, thanks to the diversification according to the document structure.

- With wrong structured web pages the method obtains bad results since the calculation of fractal dimension doesn`t give information in these cases.
Advantages

- The model takes into account the structure of the document.

- **Inheritance of values.** The importance of a parent node is distributed between all its child nodes according to the salient features using the propagation formula. Their also inherits the quota from its parent.
Conclusions

- We present the new concept of fractal dimension of text documents.
- We modify the fractal summarization method, removing the shortcomings, and using the fractal dimension of text documents getting good results.
- Experiments show the good performance of the method in only few seconds, for example in the web.
- **In the future,**
  - We want to use a similarity measure taking into account the semantic of words giving a more complete solution.
  - Adapt the method when summarizing a group of documents with similar contents.
Thank you very much

Any questions?