The goal of this study is to discover a lot of serendipitous information on the Web. How to measure serendipity is a hot topic in information filtering technologies, including recommender systems. We propose a novel method for measuring serendipity, based on the Wikipedia category system. For example, a serendipitous topic “OpenCola” can be discovered, based on the clue that two categories (“cola” and “open source”) far from each other are assigned to “OpenCola”. This study examined two machine learning techniques (SVMs and regression) for judging each topic is serendipitous or not, and found that the regression-based method is appropriate for this task.

Web search engines are suitable for discovering relevant information, but... unexpected information often leads to a discovery! = serendipity

How to measure serendipity?

Features and Training Data

Triples (a topic, and two parent categories A and B) are input objects
Features
- The number of child articles in Category A and Category B describes how each category is general
- Hierarchical level of Category A and B the number of hops between the category and the top category = describes how each category is general
- The number of common articles between Category A and Category B describes the scarcity of the triple
- The distance between Category A and Category B describes how the two categories are conceptually far from each other

Training Data
- Wikipedia Japanese edition
- Randomly selected triples are annotated by a human (positive labels and negative labels)

Discovering by Classification (SVMs)

- linear kernel, polynomial kernel (quadratic and cubic), and Gaussian kernel
- R implementation kernlib is used
- 4-fold cross validation

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear kernel</td>
<td>72.88%</td>
<td>54.43%</td>
<td>62.31%</td>
</tr>
<tr>
<td>Quadratic kernel</td>
<td>69.64%</td>
<td>57.14%</td>
<td>62.86%</td>
</tr>
<tr>
<td>Cubic kernel</td>
<td>73.33%</td>
<td>55.00%</td>
<td>62.86%</td>
</tr>
<tr>
<td>Gaussian kernel</td>
<td>76.36%</td>
<td>51.85%</td>
<td>61.76%</td>
</tr>
</tbody>
</table>

(positive:negative= 200:200)

The more negative data, the less accuracy...
Our idea to extract serendipitous information from Wikipedia

Our hypothesis: The Wikipedia category system reflects how each topic is serendipitous

- **High serendipity**
  - Two categories are far from each other
    - The topic “OpenCola” is a relation which connects separate concepts “cola” and “open source”
  - The scarcity of the relation is high
    - There is only one topic which is connected to the both categories

- **Low serendipity**
  - Two categories are near from each other
    - The topic “A&W Cream Soda” is a relation which connects neighbor concepts “cream soda” and “soft drinks”
  - The scarcity of the relation is low
    - There are a lot of topics which are connected to the both categories

Extract serendipitous topics using machine learning techniques

**Recommended systems**
- Indication that unexpectedness and novelty should be used for evaluation figures (Swearingen 2001, Murakami 2008)
- Systems which use unexpectedness and novelty have high degree of satisfaction (Mcenee 2006, Murakami 2009)

**Contents hole search**
- Discovering “contents holes”, i.e. topic not mentioned in communities (e.g. BBS, SNS), using Wikipedia (Nadamoto 2009)

**Discovering by Regression**
- Non-linear regression (cubic)
  - Add one dimension for scores (1 for positive, 0 for negative)
  - The same training data as SVMs are used
  - Topics related to two categories “Michael Jackson” (44 topics) and “global warming” (64 topics) are tested, and evaluated by a human

**11-points average precision**

<table>
<thead>
<tr>
<th>CategoryA</th>
<th>CategoryB</th>
<th>Article</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGBTの人物 (LGBT person)</td>
<td>マイケル・ジャクソン (Michael Jackson)</td>
<td>リサ・プレリー・プレスリー (Lisa Marie Presley)</td>
<td>10</td>
</tr>
<tr>
<td>マイケル・ジャクソン (Michael Jackson)</td>
<td>著名な動物 (Famous animals)</td>
<td>ハッブル (Hubble)</td>
<td>3</td>
</tr>
<tr>
<td>CategoryA</td>
<td>CategoryB</td>
<td>Article</td>
<td>Rank</td>
</tr>
<tr>
<td>地球温暖化 (Global warming)</td>
<td>ワールド・ジャンプ (World Jump Day)</td>
<td>ドキュメンタリー番組 (Documentary television films)</td>
<td>1</td>
</tr>
<tr>
<td>地球温暖化 (Global warming)</td>
<td>ドキュメンタリー番組 (Documentary television films)</td>
<td>地球温暖化詐欺 (The Great Global Warming Scindle)</td>
<td>9</td>
</tr>
</tbody>
</table>

**Conclusion and Future Work**

**Conclusion**
- A novel approach to discover serendipitous topics from Wikipedia
  - The Wikipedia category system works well
  - The Regression-based method is suitable for this task

**Future Work**
- Web-based evaluation
- Relearning by acquiring a large number of labeled data by users’ feedback