# Image Searching and Browsing by Active Aspect-Based Relevance Learning

Mark J. Huiskes

Centre for Mathematics and Computer Science (CWI),
Amsterdam, The Netherlands
Mark.Huiskes@cwi.nl

**Abstract.** Aspect-based relevance learning is a relevance feedback scheme based on a natural model of relevance in terms of image aspects. In this paper we propose a number of active learning and interaction strategies, capitalizing on the transparency of the aspect-based framework. Additionally, we demonstrate that, relative to other schemes, aspect-based relevance learning upholds its retrieval performance well under feedback consisting mainly of example images that are only partially relevant.

## 1 Introduction

For both image and video retrieval, relevance feedback has become a key process in iteratively adapting retrieval results to the user's wishes; [1] and [2] give recent reviews of the state-of-the-art of CBIR relevance feedback. In modern retrieval systems we are increasingly working with large numbers of heterogeneous features. Image descriptions now consist of a combination of low-level features (often various similarity spaces), of segmentation-based features, of learned classifications, and of high-level, possibly context-derived, metadata. The problem that relevance feedback schemes must be able to deal with only few samples in very high-dimensional feature space is thus ever more urgent.

The main inference task is to figure out which feature values, or feature value combinations, lead to high relevance. This essentially comes down to analyzing the clustering behavior of the example images provided in the relevance feedback cycles. However, as we have argued in [3], generalizing from the clustering of the small number of available example images faces a number of serious challenges. First, there is an issue of misleading clustering. Typically there will be many features that are not relevant but this does not mean that examples will not cluster for such features: often they will, namely at feature values with high prior probability. Additionally, selected examples are in practice often only *partially relevant*, especially in the critical initial stages of a searching or browsing session. This means that for the features that are relevant to the user, examples often cluster only to a limited extent. We have shown that for these reasons much can be gained by taking into account database feature value distributions, as this allows us to assess the significance of example clustering. In [3] we proposed aspect-based relevance learning, in which aspects are defined as feature predicates serving as natural units of relevance, for which indeed the significance

H. Sundaram et al. (Eds.): CIVR 2006, LNCS 4071, pp. 211–220, 2006.

of associated example clustering can be quantified. This provides a principled method to make sure that feature value regions are only selected as relevant if evidence is sufficiently strong to support the hypothesis that clustering there is not occurring by chance. In section 2 we briefly review this selection procedure, as well as the construction of aspects.

This article has two main contributions. First, in section 3, we extend aspect-based relevance learning with a number of active learning and interaction strategies. Active learning is an increasingly studied topic in image retrieval (e.g. [4,5,6,7]) which aims to reduce the number of images that need to be labeled by the user by presenting him with images that are particularly informative to the system. We propose a method for selecting such images in the aspect-based framework. Furthermore, we discuss interaction strategies which aim to provide the user with a set of tools to (i) obtain a clear insight in the system's inference process, and (ii) improve aspect data during retrieval sessions. All strategies have been implemented in the "Aspect Explorer" retrieval system.

Second, in section 4, we present an experimental study analyzing how well various relevance schemes uphold their retrieval performance under feedback consisting mainly of example images that are only partially relevant. We compare the performance of aspect-based relevance learning to various feature re-weighting schemes as well as to SVM-based methods. Tests are performed using a large commercial database of decoration designs for which a wide variety of low-level and high-level features have been computed.

## 2 Aspect-Based Relevance Learning

In the following, we assume feedback example selection is implemented by presenting images in a clickable selection display, consisting of a grid of thumbnail images. The number of images inspected per cycle may be larger as the user can leaf through the selection displays. The examples and counterexamples, selected by the user as feedback, are collected in positive and negative *example sets*. At each cycle of the feedback process the user updates the examples in the example sets by either: (i) selecting new images as positive or negative examples, adding them to their respective sets; (ii) removing images from the sets, i.e. the sets are preserved from cycle to cycle unless images are no longer deemed representative enough and are deleted. Based on the example sets a new relevance ranking is determined for display in the next feedback cycle.

### 2.1 Aspect Selection

In [3] we treat images as sets of aspects, where we understand an "aspect" simply as a property which an image either has or has not, and for which we intend to resolve its effect on perceived relevance as a unit. Aspects can thus be explicitly defined in terms of conditions on feature values, i.e. as derived binary features that model a specific perceptual quality, but can also live solely in the "eye of the beholder". In aspect-based relevance learning the feedback data available at the end of each cycle is used foremost to establish the relevance effect (neutral, positive or negative) of the various aspects.

The main idea is that we must find those aspects for which the user has *actively* selected more examples with that aspect than may be expected to arise by chance only. To this end, we first define the *aspect image frequency* $p_{\mathrm{db}}$ as the fraction of images in the database having a certain aspect. Given this frequency, we can model the probability distribution of the number of examples having an aspect when that aspect is neutral, viz. by assuming that it will approximately follow the distribution of aspect occurrence in the database.

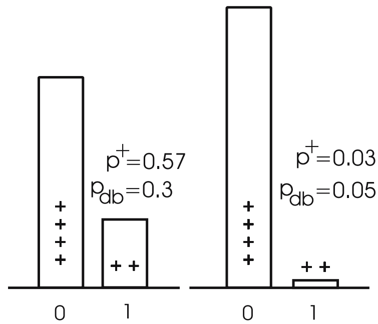



**Fig. 1.** The significance of example clustering depends on aspect image frequency. Two aspects are shown with bars representing the prior distribution of aspect possession (corresponding to the indicated $p_{\mathrm{db}}$ values). For the second aspect the clustering of the two examples with the aspect is more significant than for the first. Note how the method deals with partial relevance: clustering may be significant even if the majority of examples do not have an aspect.

In the following we analyze the positive example set; the negative examples are treated analogously. Let $n^+$ be the current total number of positive examples, and $N^+$ the number of positive examples that possess an aspect. For each aspect, we consider an *independence hypothesis* $H_0^+$, stating that the aspect is neutral to the user. Under this hypothesis we model aspect possession of an example image as a Bernoulli variable with probability $p_{\mathrm{db}}$; consequently, the number of positives with given aspect can be modeled as a binomial variable with parameters $n^+$ and $p_{\mathrm{db}}$.

An aspect is selected as positive only if the independence hypothesis can be rejected based on an unexpectedly high number of example images with that aspect. The $p$-value associated with the hypothesis is thus the probability of finding $N^+$ or more images with aspect in a set of $n^+$ random images

$$p^+(N^+) = \sum_{i=N^+}^{n^+} \binom{n^+}{i} p_{\mathrm{db}}^i (1 - p_{\mathrm{db}})^{(n^+ - i)}. \tag{1}$$

We select only those aspects for which their $p^+$-values are below a certain threshold $p_0^+$. As Fig. 1 shows, this approach has the benefit that feature selection is not just based on the clustering behavior of the examples. Rather, it analyzes this clustering in its context, viz. the current database. In addition, by taking into account feature value distributions, we are not dependent on negative examples to down-weight positives that cluster at aspects with low saliency. This means negatives can be used to indicate which aspects are not desired, but are not required for the sole purpose of getting sufficient data for classification.

### 2.2 Aspect Construction

We can divide the methods for constructing aspects from features into supervised and unsupervised types. Unsupervised methods are, for instance, automatic quantization of one-dimensional features and unsupervised clustering methods for higher dimensional feature spaces. Supervised methods may use general classification methods such as SVMs, boosting or prototype-based methods to generate aspects based on sets of annotated sample images. Another supervised method is supervised quantization, by which the user manually selects feature value ranges of interest as aspects. Though unsupervised methods require less work, in general we prefer supervised methods as they tend to lead to aspects that are more perceptually meaningful. Note that binary or discrete features can be converted directly into aspects. For more details, see [3].

## 3 Active Learning and Interaction

Figure 2 shows the Aspect Explorer retrieval interface.

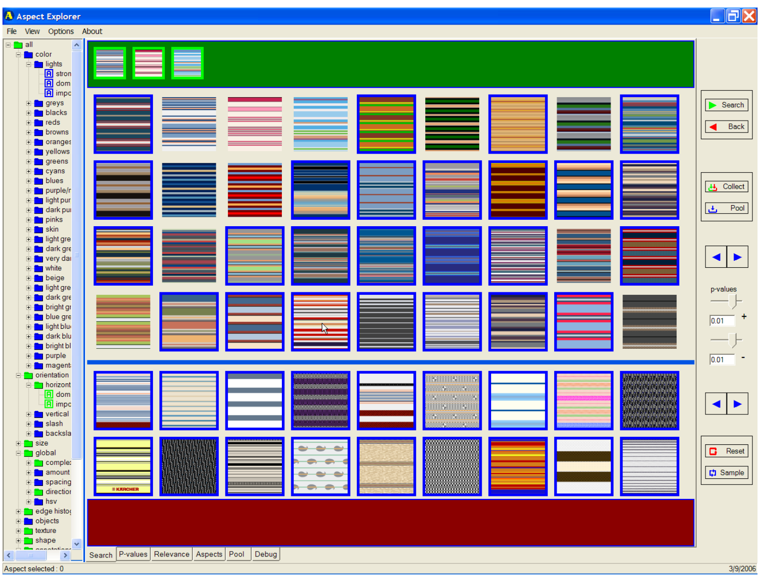



**Fig. 2.** Aspect Explorer retrieval interface. Selected aspects are highlighted in green (positive) and red (negative; not shown) in the aspect hierarchy on the left. The positive example set is in the green section at the top; negatives in the red section at the bottom. The top part of the selection display, the "relevance display" (RD) shows most relevant images; the lower "information display" (ID) shows images deemed most informative to the system; both displays can be leafed through to browse lower-ranked images.

### 3.1 Relevance Ranking

The Aspect Explorer system implements the aspect selection scheme detailed above. The main interaction takes place by selection of examples in the "relevance display" (RD). The images displayed there have been determined directly from the selected aspects through a greedy ranking algorithm. Let $A^+$ and $A^-$ be the index sets of accepted positive and negative aspects, respectively, then the relevance $\text{rel}_i$ for image $i$ is defined as $\text{rel}_i = \sum_{j \in A^+} M_{ij} - \sum_{j \in (A^- / A^+)} M_{ij}$. Here $M$ is the aspect matrix with boolean variables $M_{ij}$ flagging possession of aspect $j$ for image $i$.

### 3.2 Active Learning

The RD shows images inferred to be most relevant to the user given the current example sets. However, these images are not necessarily most informative to the system for improving its retrieval results for the next cycle. Active learning deals with methods to determine images, to be presented to the user for labeling, which allow the system to learn fast. Since, in aspect-based relevance learning, feedback data is used primarily for aspect selection, our aim is to find images that may resolve the status of the most uncertain aspects. The resulting "informative" images will be presented to the user in the "information display" (ID), see Fig. 2. As both selection displays are accessible simultaneously, the ID is designed to show images complementary to images in the RD.

The method proposed here constructs a ranking of images by informativeness on aspects that are uncertain in the sense that their selection $p^+$-values are close to the threshold. For reasons described below, we focus on aspects with $p$-values above the threshold, i.e. which have not yet been selected.

An aspect that is not yet selected, but shows promise given its $p$-value, may be selected in the next cycle if a user selects one or more images with that aspect as positive example. However, simply presenting the user with images having a particular uncertain aspect is not useful as their appearance is generally dominated to a large extent by aspects the user is not interested in. We found that this may be resolved by presenting images that have the uncertain aspect, as well as a relatively high relevance score. Closely related to this, we found that it is important to take into account the *aspect accumulation* of the uncertain aspects. The accumulation is defined as the fraction of images having the particular aspect in the, say 50, top-ranking relevant images. It turns out that it is fairly common for uncertain aspects to already be highly accumulated. The reason is that uncertain aspects often show a correlation with selected aspects. Such accumulated aspects are not interesting for constructing informative images as the user already has sufficient access to such images in the RD.

We conclude that we must present informative images for uncertain aspects that have not yet accumulated in the top-ranking images. To this end, we first sort aspects according to an information score consisting of a weighted combination of the proximity of their $p^+$-value to the threshold and their accumulation fraction. We found that taking the top-5 of lowest scores provides a suitable set of uncertain aspects. Next, for each uncertain aspect we rank the images having that aspect by relevance. The final information ranking is constructed by interleaving the rankings of the uncertain aspects. Images with a high information ranking may still occur in the top images of the RD. These are not repeated in the ID but rather highlighted in the RD (in blue).

Experience with the Aspect Explorer system shows that the informative images, thus constructed, often provide interesting example images as they tend to have promising aspects that are not yet available in the RD. The ID ranking can thus be viewed as a less greedy ranking, for which more than just the selected aspects have been taken into account. An example is shown in Fig. 2, where in the example set there is some evidence that the user may be interested in light colors. In the ID several relevant images with light colors are indeed suggested.

### 3.3 Advanced Interaction

For interaction scenarios where the emphasis is on ease of use of the interface, e.g. letting customers browse a design collection in a shop, restricting the interface mainly to selection of positive and negative examples from the relevance display (RD) is probably the best choice. For expert users, we can use more advanced features such as the ID, as well as provide insight into the workings of the system. We also discuss a number of interaction strategies that allow the user to correct the system, both at the aspect selection level and the image aspect data level. The latter means that an authorized user may correct aspect data if this proofs merited during the course of a retrieval session.

An important interface component, contributing directly to the transparency of the inference process, is the aspect hierarchy tree, see Fig. 2; its structure is described by a simple XML-file, providing aspect names and their grouping relations. For example, all color aspects are organized in a color group which contains subgroups for specific colors, which, in turn, contain the aspects modeling the various levels of importance of that color. The groups serve to organize aspects and to allow modification of the status of many aspects simultaneously. The groups may be aspects themselves, e.g. the aspect “flower” is also a group containing aspects such as “rose” and “tulip”.

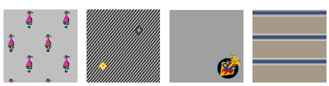
The aspect tree is used to summarize the state of the inference engine at the end of a feedback cycle in terms of the selected positive and negative aspects, and to allow the user to give direct feedback on this state. Selected positive aspects are highlighted in green, negatives in red. If desired, the user can provide feedback directly in terms of the aspects, both through *soft* and *hard* selection. Soft selection means the aspect is treated as if the selection had occurred from passing the selection hypothesis test. Hard selection is stronger in that possession of the selected aspect becomes binding, restricting the search to a subset of database images having (or not having) the particular aspect. It is also possible to manually “neutralize” aspects selected by the system.

As mentioned, the interaction design has been aimed at making it possible to correct aspect values, whenever a situation is encountered during the retrieval process which shows that a value is in error. We discuss two scenarios. As a first one, consider the situation that the evaluation of a certain aspect as positive, negative, or neutral is surprising to a user. The user may then click this aspect in the aspect tree. This will then highlight all images in the RD, ID and example sets based on the possession of that aspect: green if it has the aspect, red if it does not. Authorized users may then proceed to edit the aspect values in various ways, e.g. by clicking an example image to toggle its value. Other methods use image dragging or pop-up menus to move or copy the image to a different aspect. In the second scenario the user finds an image with an unexpectedly high ranking (to him). In such case, he may inspect its possession of selected aspects, either in a dialog window or the aspect tree; and, again, edit if necessary.

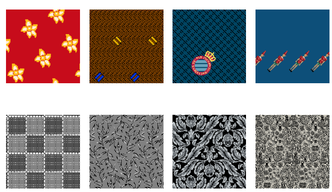
A final useful interaction type is directed at improving aspects in batch-mode. To this end, a separate tab is used to show a selection display divided in two parts: one for images with the selected aspect, one for without; both may be leafed through. This facilitates easy improvement of aspect data by dragging images from one side to the other.

## 4 Query Simulation Experiments

The main job of a relevance feedback scheme is to produce a relevance ranking of the database images, based on an analysis of the example sets obtained at the end of a feedback cycle. To compare the performance of different schemes we consider the quality of the rankings produced. We will do so for queries represented by simulated example sets. The example sets are simulated in such a way that it is clear what represents a good ranking. Performance is quantified by means of precision-recall graphs for the target images.



(a)



(b)

**Fig. 3.** Two simulated example sets, for target aspects "background, large" and "grey, dominant", generated according to the (a) full relevance scenario, and (b) partial relevance scenario

We first select a number of *target aspects*. Each such aspect represents a feature value, or range of feature values, the user is actively interested in. The target aspects are randomly sampled from aspects with a $p_{\mathrm{db}}$ value below a given threshold. This assures that the target aspects are saliently present and perceptually relevant. In this study we consider only positive aspects; we do not simulate negative aspects. The target aspects directly correspond to a set of *target images*, viz. those images in the database that possess all target aspects.

We consider two main scenarios for generating the positive example sets. In the *full relevance* scenario we simply sample a subset of the target image set. Fig. 3 (a) shows a generated example set of 4 target images for two target aspects "background, large" and "grey, dominant".

The second scenario is intended to test how well the relevance feedback schemes can deal with example sets consisting of images that are only partially relevant. In this *partial relevance* scenario we generate a fixed number of images for each of the target aspects. The images are randomly sampled from those images that have one target aspect, but do not have the remaining target aspects (or, as few as possible, if no such images exist). Fig. 3 (b) shows such an example set for the same two aspects as before. Observe, for instance, that the last 4 images are grey, but have relatively little background. Compared to the first scenario, where the simulated example sets consist strictly of fully relevant images, this scenario is at the other end of the feedback spectrum, in the sense that feedback is provided here exclusively by means of partially relevant images. In practice, example sets will usually be a mix of these two extreme scenarios. For both scenarios, we test both with and without additional simulated negative examples. Negative examples are randomly sampled from the images that have none of the target aspects. This is mainly for the benefit of the other learning methods as for the aspect-based method they are not needed.

We compared the aspect-based relevance learning method (ARL) to a number of other methods. Two are based on support vector machines, both implemented using [8]. The first (1-SVM) is the one-class SVM method of [9]. The second is a standard $\nu$-SVM method for classifying positive and negative examples, see for instance [10]. We tested both a Gaussian (2-SVM(g)) and a polynomial (2-SVM(p)) kernel. Additionally, we considered feature re-weighting methods with relevance ranking based on the moving query point mechanism ([11]). The relevance ranking follows from the sorted distances of the images to the moving query point, using two methods to determine the feature weights used in the distance measure. The first (FRW1) uses feature weights that are inversely proportional to the example feature value variance. The second (FRW2), proposed in [12], uses a more advanced weighting scheme, particularly when also negative examples are available. For ARL we took $p = 0.01$ as the $p$-value selection threshold, which is also the default threshold in Aspect Explorer.

Testing is based on aspects (ARL) and features (the other methods) for a large commercial database of decoration designs. To characterize decoration designs we have selected and developed a variety of features suitable for representing both their global appearance (e.g. color, texture, complexity), as well as their elements (shape, size, number, variation, spatial organization). Metadata descriptions in terms of a hierarchy of 42 keywords (e.g."geometric", "flower") are also used. A detailed description of all features is provided in [13]. Features were computed for 7078 images. From the 161 available features, a total of 585 aspects were derived.

### 4.1 Experimental Results

Fig. 4 shows precision-recall graphs based on the ranking of the target images for the scenarios outlined above. For each run, 1000 simulations were performed. For the simulations, random target aspects were sampled such that the number of target images was always at least 10, and additionally, the selected aspects corresponded to different features. Only salient aspects with $p_{\mathrm{db}}$ value of at most 0.1 were considered as target aspects, and for each aspect 4 example images were selected. Several variations to the experiments have been performed (e.g. taking more example images, increasing the saliency threshold, using a different minimum number of target images, or leaving out various feature groups). These showed roughly the same relative performance between ARL and the other methods.

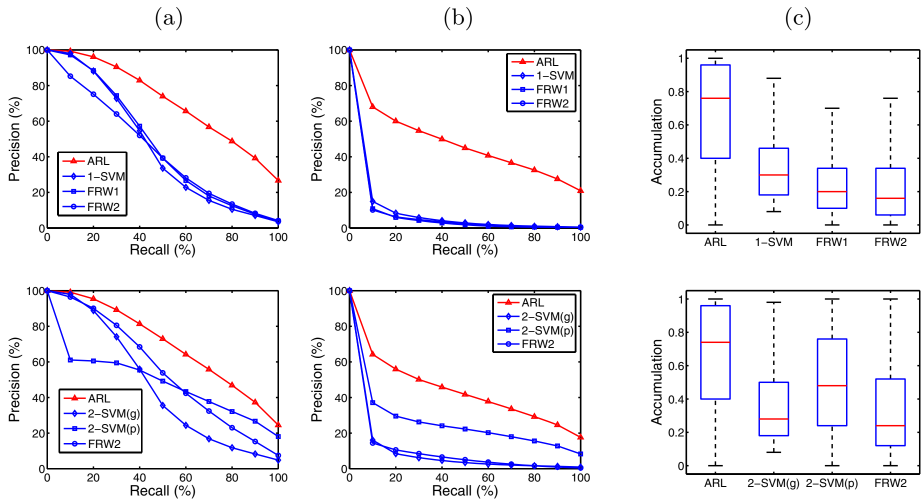



**Fig. 4.** Target image precision-recall graphs for (a) full relevance scenario, (b) partial relevance scenario; (c) accumulation box-plots for the partial relevance scenario. The top graphs show results for just positive examples; the bottom graphs for both positive and negative examples.

In the full relevance scenario the aspect-based relevance feedback method outperforms the feature re-weighting methods mainly because the re-weighting methods assign too many features with high weights, which leads to poor precision. In the partial relevance scenario the performance of the feature re-weighting methods further deteriorates due to additional difficulty in selecting the correct features as clustering is less clear due to the partial relevance of the examples. For the SVM methods it is interesting to notice that using a polynomial kernel provides a much better performance in the partial relevance scenario than the Gaussian kernel. Fig. 4 (c) demonstrates that ARL leads to a higher accumulation of target aspects in the top ranking images. Shown are the average accumulation fractions (defined in section 3.2) over the target aspects.

## 5 Conclusion

We have demonstrated that, relative to other schemes, aspect-based relevance learning upholds its retrieval performance well under feedback consisting mainly of example images that are only partially relevant. As a result, the aspect-based approach leads to a natural interaction where generally "what you click is what you get". When this is not the case this is often caused by mistakes in the features or aspects. Several interaction strategies have been presented to correct such mistakes on the fly. Finally, presenting the user with images based on an active learning method was found to provide a useful extension to presenting most relevant images. Such images can resolve uncertainty on whether an aspect should be selected and, by their construction, often have aspects interesting to the user not yet available in the RD.

## Acknowledgement

This research was partially supported by EC FP6 NoE MUSCLE (FP6-507752).

## References

1. Zhang, H., Zheng, C., Li, M., Su, Z.: Relevance feedback and learning in content-based image search.WWW: Internet and web information systems **6** (2003) 131–155
2. Zhou, X., Huang, T.: Relevance feedback in image retrieval: a comprehensive review. ACM Multimedia Systems Journal **8**(6) (2003) 536–544
3. Huiskes, M.: Aspect-based relevance learning for image retrieval. In Leow, W., ed.: Proceedings of CIVR05, LNCS 3568. Springer (2005) 639–649
4. Tong, S., Chang, E.: Support vector machine active learning for image retrieval. Proc. of 9th ACM Int. Conference on Multimedia (2001) 107–118
5. Cord, M., Gosselin, P., Philipp-Foliguet, S.: Stochastic exploration and active learning for image retrieval. Image and Vis. Computing (Acc. for publ., Jan 2006)
6. Zhou, Z., Chen, K., Jiang, Y.: Exploiting unlabeled data in content-based image retrieval. ECML'04, LNAI 3201 **8**(6) (2004) 525–536
7. Zhang, H., Su, Z.: Relevance feedback in CBIR. In Zhou, X., Pu, P., eds.: Visual and Multimedia Information Systems. Kluwer Academic Publishers (2002) 21–35
8. Canu, S., Grandvalet, Y., Guigue, V., Rakotomamonjy, A.: SVM and kernel methods Matlab toolbox. Perception Syst. et Inf., INSA de Rouen, France (2005)
9. Chen, Y., Zhou, X., Huang, T.: One-class SVM for learning in image retrieval. Proc. IEEE ICIP 2001, Thessaloniki, Greece **1** (2001) 34–37
10. Hoi, C., Lyu, M.: Biased support vector machine for relevance feedback in image retrieval. Proc. Intl. Joint Conf. on Neural Networks, Budapest, Hungary (2004)
11. Rocchio Jr., J.: Relevance feedback in information retrieval. In Salton, G., ed.: The SMART retrieval system: experiments in automatic document processing. Prentice-Hall (1971) 313–323
12. Ciocca, G., Schettini, R.: A relevance feedback mechanism for content-based image retrieval. Information Processing and Management **35**(5) (1999) 605–632
13. Huiskes, M., Pauwels, E.: Indexing, learning and CBR for special purpose image databases. In Zelkowitz, M., ed.: Advances in Computers. Elsevier (2005)