Integrating and Querying Parallel Leaf Shape Descriptions

Shenghui Wang[1] and Jeff Z. Pan[2]

[1] School of Computer Science, University of Manchester, UK
[2] Department of Computing Science, University of Aberdeen, UK

Abstract. Information integration and retrieval have been important problems for many information systems — it is hard to combine new information with any other piece of related information we already possess, and to make them both available for application queries. Many ontology-based applications are still cautious about integrating and retrieving information from natural language (NL) documents, preferring structured or semi-structured sources. In this paper, we investigate how to use ontologies to facilitate integrating and querying information on parallel leaf shape descriptions from NL documents. Our approach takes advantage of ontologies to precisely represent the semantics in shape description, to integrates parallel descriptions according to their semantic distances, and to answer shape-related species identification queries. From this highly specialised domain, we learn a set of more general methodological rules, which could be useful in other domains.

1 Introduction

Information integration and retrieval have been important problems for many information systems [1] — it is hard to combine new information with any other piece of related information we already possess, and to make them both available for application queries. Most information in *descriptive* domains is only available in natural language (NL) form and often comes *parallel*, *i.e.*, the same objects or phenomena are described in multiple free-styled documents [2]. With ontologies being shared understandings of application domains, ontology-based integration and retrieval [3] is a promising direction. However, many ontology-based applications avoiding integrating and retrieving information from NL documents, preferring structured or semi-structured sources, such as databases and XML documents.

In this paper, we investigate how to use ontologies to facilitate integrating and querying information on parallel leaf shape descriptions from botanical documents. As one of the premier descriptive sciences, botany offers a wealth of material on which to test our methods. Our observation is that if the parallel information can be extracted and represented in a uniform ontology, the explicitly written information can be accessed easily and the implicit knowledge can also be deduced naturally by applying reasoning on the whole ontology. We have recently demonstrated that it is feasible for an ontology-based system to use this method to capture, represent and use the semantics of colour descriptions from

I. Cruz et al. (Eds.): ISWC 2006, LNCS 4273, pp. 668–681, 2006.

botanical documents [4]. In this paper, we focus on another specialised aspect — leaf shape descriptions.

As a highly domain-dependent property, shapes are not easily described in NL. Unlike colours, a specialist terminology is used to describe shapes that naturally occur in each domain, combined with general NL syntax. For instance, the leaves of the aspen trees are described differently in five floras:[1]

- broadly ovate to suborbicular or oblate-orbicular
- broadly ovate to orbicular
- kidney-shaped, reniform or oblate
- suborbicular
- almost round

To capture the semantics in these descriptions and formalise them into an ontology system is our concern. Our approach takes advantage of ontologies to represent the semantics in shape descriptions precisely, to integrate parallel descriptions according to their semantic distances, and to answer shape-related species identification queries.

1. Firstly, we need an appropriate semantic model in which the semantics in shape descriptions can be captured and the compatibility between descriptions can be measured. We adopt a known shape model, called SuperFormula [5], to model common leaf shape terms. Based on this we derive a domain-dependent four-feature leaf shape model. The semantics of complex descriptions are precisely constructed from those of simple terms by applying a small number of morpho-syntactic rules. The quantitative semantics is then represented in the OWL-Eu ontology language [6].
2. Secondly, we propose a distance function, based on the four-feature leaf shape model, to calculate distances between parallel information (*e.g.*, the distance between "linear to ovate" and "narrowly elliptic"), so as to facilitate a proper strategy of integrating such information.
3. Thirdly, we use the OWL-Eu subsumption reasoning to check if one shape description is more general than another one. Such a reasoning service is helpful in answering species identification queries, for example, to search all species which have "ovate to elliptic" leaves (more examples in Section 5) over the integrated information.

In order to check the feasibility of the above approach, we develop and implement a shape reasoner, based on the FaCT-DG Description Logic reasoner [7,8]. The shape reasoner integrates parallel shape information based on their semantic distances; it also answers queries over the integrated information. We will show that semantic distances can also improve the presentation of the query results: they help by (i) measuring how well the results match the query, and (ii) presenting the best results first. We evaluate our approach in two steps. Firstly, we ask a domain expert to check how good our proposed semantic model and

[1] A flora is a treatise on or list of the plants of an area or a period.

semantic distance function are. Secondly, we evaluate the query results from our shape reasoner based on the reliable semantic distance function.

The rest of the paper is structured as follows. Section 2 introduces a known shape model and our four-feature leaf shape model. In Section 3, we show how the semantics in a complex leaf shape description is constructed and represented formally. Section 4 introduces distance-based integration and some experimental results. Section 5 investigates how to query on the integrated information and improve the presentation of returned results by ranking them, based on their degree of match to a particular query. Section 6 discusses related work and Section 7 concludes this paper.

2 A Multi-parametric Semantic Model for Leaf Shapes

Shape modelling is not easy, in the sense that it is highly domain dependent. People have tried to use cylinders [9] or superquadrics [10] as primitives to model abstract shapes. For real shapes in nature, several modelling methods have also been tried, such as interpolation methods which use polynomials or splines to fit curves. Since the pioneering work of D'Arcy Thompson [11], biomathematicians have investigated describing natural shapes and forms by using morphometric methods [12]. Outlines and landmark-based patterns are used to represent natural shapes. However, their high-dimensional representation cannot be interpreted easily and is not suitable in a logic-based system.

Gielis [5] recently proposed the Superformula, which in polar co-ordinates (r, θ), is:

$$r(\theta) = \frac{1}{\sqrt[n_1]{(|\frac{1}{a}\cos(\frac{m}{4}\theta)|)^{n_2} + (|\frac{1}{b}\sin(\frac{m}{4}\theta)|)^{n_3}}} \tag{1}$$

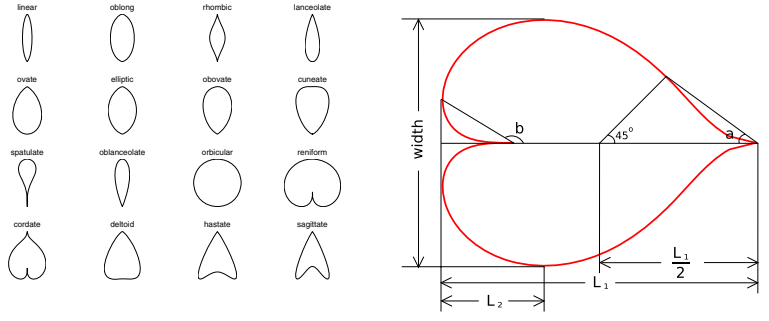
This can generate approximations to many naturally occurring shapes and forms. Although it is not easy to find the precise parameters (m, a, b, n_1, n_2, n_3) for a particular shape, the simplicity and expressiveness of this formula encouraged us to use it for modelling leaf shapes.

Here, we consider only *simple* leaves[2] for demonstrating the feasibility of our method. We selected 21 common simple leaf shape terms from *Botanical Latin* [13]. Based on our experiments and experts' evaluation, for each term, we found a 6D vector (m, a, b, n_1, n_2, n_3) which generates its prototypical shape. For instance, the parameters $(2, 1, 1, 1, -0.5, 0.5)$ generates a "cordate" shape. Figure 1 (a) shows some other shapes.

The terminology is limited while real shape variations are continuous. Therefore, in order to describe continuous shape variations, one has to compare the real shapes with prototypical ones. If a real shape S_i is *similar* enough to the prototypical shape S of a shape term T, *i.e.*, their distance $d(S_i, S) < \epsilon$, then it can be named by term T. Thus each term does not correspond to a point but a region around that point in the multi-parametric space.[3] Complex leaf shape

[2] Simple leaves are entire (without teeth or lobes) and bilaterally symmetric about their main vein.

[3] According to the conceptual space theory [14], this region must be convex.



(a) Common leaf shapes generated by the SuperFormula

(b) 4-feature leaf shape model derived from the SuperFormula

Fig. 1. Leaf shape modelling

descriptions, such as "narrowly ovate to elliptic", also correspond to certain regions. Since the shape of such regions is still unknown [14], we use a simple definition: the region for a shape term contains all points whose distance to its prototype point is smaller than a predefined threshold.

Unfortunately, the six parameters of the Superformula are not directly related to any visible leaf features, which makes it extremely difficult to measure shape distance directly based on the 6D vectors. Therefore, we devised a special leaf shape model. Four basic features are calculated from the shape generated by the Superformula, see Figure 1 (b):

- length-width ratio: $f_1 = \frac{L_1}{width}$;
- the position of the widest part: $f_2 = \frac{L_2}{L_1}$;
- the apex angle: $f_3 = a$;
- the base angle: $f_4 = b$.

In this four-feature shape model, each term corresponds to a region with a small range in each feature while the region of a complex shape description is constructed from those of simple terms (see next section for details). The distance function between shape regions is defined in Section 4.

3 From NL Phrases to Ontological Representation

3.1 Morpho-syntactic Rules

One term is usually not enough to cover the natural shape variations of one species, hence complex descriptions have to be used (as shown in Section 1). In order to capture the semantics of these descriptions, we need to know how they are constructed from basic terms. We carried out a morpho-syntactic analysis

Table 1. Leaf shape description patterns

Leaf Shape Description Pattern	Example
1. Single term	"ovate"
2. Modified term	"broadly elliptic"
3. Hyphenated expression	"linear-lanceolate"
4. Range built by "to"	"oblong to elliptic"
5. Multiple ranges connected by coordinators ("and","or"), or punctuations	"linear, lanceolate or narrowly elliptic" "ovate and cordate"

on 362 leaf shape descriptions of 291 species from five floras.[4] The description patterns are summarised in Table 1.

3.2 Semantics for Complex Descriptions

The semantics of complex descriptions is constructed by applying certain operations on that of basic terms. Firstly, basic shape regions are generated, including:

Single term: Given the 6D vector of a simple term, we calculate its four features (f_1, f_2, f_3, f_4), then we generate a region with a small range in each feature, *i.e.*, $(r_{f_1}, r_{f_2}, r_{f_3}, r_{f_4})$, where $r_{f_i} = [f_i \times 0.9, f_i \times 1.1]$, for $i = 1, \ldots, 4$.

Modified term: Leaf shapes are normally modified in terms of their length-width ratio, *e.g.*, "narrowly" and "broadly." As side effects, apex and base angle also change. According to our experiments, if "narrowly" and "broadly" are defined as:

$$\begin{aligned} &\text{"narrowly:"} \quad f_1{}' = f_1 \times 1.2 \\ &\qquad\qquad\quad f_i{}' = f_i \times 0.9, \text{ for } i = 3, 4 \\ &\text{"broadly:"} \quad f_1{}' = f_1 \times 0.8 \\ &\qquad\qquad\quad f_i{}' = f_i \times 1.1, \text{ for } i = 3, 4 \end{aligned}$$

then the region around the new point $(f_1{}', f_2, f_3{}', f_4{}')$ represents the best "narrowly" and "broadly" shape of this term.

Hyphenated expression: According to the experts we consulted, a hyphenated expression "X-Y" means an intermediate shape between X and Y. The intermediate features between X and Y are calculated as follows:

$$hf_i = \frac{f_{X\,i} + f_{Y\,i}}{2}, \text{ for } \; i = 1, \ldots, 4 \tag{2}$$

The region is generated correspondingly.

Secondly, we combine basic regions to construct the region for the complex descriptions.

1. If basic shapes are connected by one or more "to"s, the final region should be the whole range from the first one to the last one. That is, the range which covers two basic regions ($r^1_{f_1}$, $r^1_{f_2}$, $r^1_{f_3}$, $r^1_{f_4}$) and ($r^2_{f_1}$, $r^2_{f_2}$, $r^2_{f_3}$, $r^2_{f_4}$) is (R_{f_1}, R_{f_2}, R_{f_3}, R_{f_4}), where $R_{f_i} = [\min(r^1_{f_i}, r^2_{f_i}), \max(r^1_{f_i}, r^2_{f_i})]$.

[4] They are *Flora of the British Isles* [15], *New Flora of the British Isles* [16], *Flora Europaea* [17], *The Wild Flower Key* [18] and *Gray's Manual of Botany* [19].

2. If basic shapes are connected by any of these symbols: "or," "and," comma (",") or slash ("/"), they are kept as separate regions, *i.e.*, disjoint from each other. Notice that "and" is treated as a disjunction symbol, because it does not indicate a logical conjunction in a NL scenario [20]. Instead, it normally indicates that the shapes could both be found in nature for the same species, similar to the meaning of "or".

By using an NL parser with corresponding operations, the semantics of a complex description can be constructed into a multi-parametric representation. Next, we need to formalise the semantics in our plant ontology.

3.3 Representing Shape Descriptions in Ontologies

As the W3C standard ontology language OWL DL [21] does not support XML Schema user-defined datatypes, we use the OWL-Eu language [6] suggested by a W3C Note [22] from the Semantic Web Best Practice and Deployment Working Group. OWL-Eu supports customised datatypes through unary datatype expressions (or simply datatype expressions) based on unary datatype groups. This support of customised datatypes is just what we need here to capture feature information of leave shapes. Like an OWL DL ontology, an OWL-Eu ontology typically contains a set of class axioms, property axioms and individual axioms.[5] Here we use the FaCT-DG ontology reasoner, a Datatype Group extension of the FaCT reasoner, which supports reasoning in OWL-Eu ontologies that do not contain nominals.[6]

The fragment of our plant ontology $\mathcal{O}_s$ contains Species, Leaf and LeafShape as primitive classes; important object properties include $hasPart$ and $hasShape$; important datatype properties include $hasLengthWidthRatio$, $hasBroadestPosition$, $hasApexAngle$ and $hasBaseAngle$, which are all *functional* properties.[7] Each datatype property and its range is also defined, for example,

$$\texttt{DatatypeProperty}(hasBaseAngle\ \texttt{Functional range(and(}\geq 0, \leq 180\texttt{)))},$$

where $\texttt{and}(\geq 0, \leq 180)$ is a unary conjunctive datatype expression representing the sub-type [0,180] of Integer. Typical relations between classes include:

$\mathsf{Species} \sqsubseteq \exists hasPart.\mathsf{Leaf}$ (Each species has a part: leaf)
$\mathsf{Leaf} \sqsubseteq \exists hasShape.\mathsf{LeafShape}$ (Each leaf has a property: leafshape)

Actural leaf shapes are defined using the above primitive classes and properties, where datatype expressions are used to restrict the values of four features. For example, the shape "ovate" is defined as the following OWL-Eu class:

$$\mathsf{Ovate} \equiv \mathsf{LeafShape} \sqcap$$
$$\exists hasLengthWidthRatio.(\texttt{and}(\geq 15, \leq 18)) \sqcap \exists hasApexAngle.(\texttt{and}(\geq 41, \leq 50))$$
$$\exists hasBroadestPosition.(\texttt{and}(\geq 39, \leq 43)) \sqcap \exists hasBaseAngle.(\texttt{and}(\geq 59, \leq 73))$$

[5] See [6] for more details on datatype expressions and unary datatype groups.
[6] Details of the FaCT-DG reasoner as well as its flexible reasoning architecture can be found in [8] and `http://www.csd.abdn.ac.uk/$\sim$jpan/factdg/`.
[7] A functional datatype property relates an object with at most one data value.

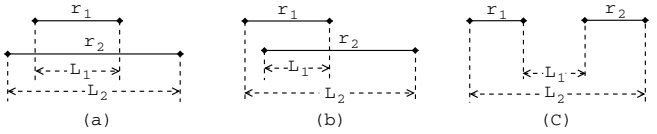


Fig. 2. Three relations between two ranges

Similarly, complex shape descriptions are also represented as OWL-Eu classes based on the regions with constraints on the four features. Ontological representations of shape descriptions enable us to carry out species identification queries based on their leaf shapes (see Section 5 for more details).

4 Distance-Based Integration

The example in Section 1 shows that parallel descriptions are very common among existing floras. In this section, we present a distance-based integration approach for parallel shape descriptions.

4.1 Distance Definition for Leaf Shape Descriptions

Parallel information is assumed to be complementary, possibly with a certain degree of overlap.[8] It is not appropriate to simply combine two or more pieces of information without carefully studying how similar or how different they are. However, measuring the distances between shape descriptions is not easy, while defining the distance between shapes itself is already an inherently ill-defined problem. For example, how far is "linear to ovate" from "linear to elliptic"?

As introduced in Section 3, a complex shape description is translated into a vector, and each element is a range in one feature, *i.e.*, $(R_{f_1}, R_{f_2}, R_{f_3}, R_{f_4})$. In order to calculate the distance between such vectors, distances in each element range should first be calculated. There are three different types of relations between two ranges, shown in Figure 2. We define the following distance function for two arbitrary ranges r_1 and r_2:

$$d(r_1, r_2) = \begin{cases} 1 - \frac{L_1}{L_2} & \text{if } r_1 \text{ and } r_2 \text{ overlap} \\ 1 + \frac{L_1}{L_2} & \text{otherwise;} \end{cases} \tag{3}$$

where L_2 is the length of minimal super-range which contains both r_1 and r_2, and L_1 is defined as follows: when r_1 and r_2 overlap (see (a) and (b)), L_1 is the length of the overlapping part; otherwise, for (c), L_1 is the length of the gap between two ranges. If two ranges r_1 and r_2 only share one point, we say they *meet* each other and $L_1 = 0$.

The distance $d(r_1, r_2)$ is nicely scaled into the range $[0, 2)$: if $d(r_1, r_2) = 0$, r_1 equals r_2; if $0 < d(r_1, r_2) < 1$, r_1 and r_2 overlap; if $d(r_1, r_2) = 1$, r_1 meets r_2; if

[8] [23] showed that, when information was collected from six parallel descriptions of a representative sample of plant species, over half the data points came from a single source, while only 2% showed outright disagreement between sources.

$1 < d(r_1, r_2) < 2$, r_1 and r_2 are disjoint; as two ranges move further apart from each other, the distance gets closer to 2.

The distance along each feature is calculated by using Formula (3). The whole distance between two shape regions R_1 and R_2 is then calculated as:

$$d(R_1, R_2) = \sum_{i=1}^{4} w_i \times d_{f_i} \tag{4}$$

where d_{f_i} is the distance in the feature f_i, w_i is the corresponding weight for the feature f_i, and $\sum_{i=1}^{4} w_i = 1$ holds.[9] The $d(R_1, R_2)$ has similar mathematical properties to $d(r_1, r_2)$, but is harder to interpret due to the influence of the weighting. According to our experiments with a domain expert from the Museum of Manchester,[10] this similarity distance function is valid and corresponds closely to how experts judge similarity between shapes.

4.2 Integration Based on Semantic Distances

We can now compute the distance between two descriptions, as calculated by Formula 4. If two descriptions are "close" or "similar" enough, although they might not be identical (for various reasons), it is better to combine them into one single "super-description" so that redundancies can be removed. Otherwise, it is safer to leave them separate because they are likely to provide complementary information of the same object. If a reasonable threshold is chosen, our integration process can automatically combine similar descriptions and keep others separate.

So, for a single species, the recursive integration process on the collections of shape regions from parallel descriptions is as follows:

Step 1 Calculate the distances between any pair of regions.

Step 2 Select two closest regions and check whether they are similar enough, *i.e.*, whether their distance is less than the threshold. If they are not similar enough then the integration stops; otherwise, the smallest region containing both of them is generated (this is same operation as building "to" ranges). This new region replaces the original two as their integrated result.

Step 3 Go back to Step 1 to check the updated collection of regions to see whether there are any further pairs of regions requiring integration.

4.3 Experiments on Integration

We selected 410 species from the floras mentioned in Section 3 and the online efloras,[11] so that each of the selected species is described in at least two flo-

[9] From our statistical analysis on real text data, f_2 is the most distinguishing feature. However, there is no satisfactory way to find the optimal weights.

[10] The contact information for our domain expert is available on request.

[11] This is an international project (http://www.efloras.org/) which collects plant taxonomy data from several main floras, such as *Flora of China*, *Flora of North America*, *Flora of Pakistan*, etc. Plant species descriptions are available in electronic form, but are still written in the common style of floras, *i.e.*, semi-NL.

Table 2. Examples of integration results, where R_{f_1} is the range of the length-width ratio, R_{f_2} is the range of the position of the widest part, R_{f_3} is the range of the apex angle: R_{f_4} is the range of the base angle

Species	Leaf Shape Descriptions	Integration Results			
		R_{f_1}	R_{f_2}	R_{f_3}	R_{f_4}
Salix pentandra (Laurel willow)	ovate or ovate-elliptical to elliptical- or obovate-lanceolate broadly lanceolate to ovate-oblong broadly elliptical broadly lanceolate, ovate-oblong, or elliptic-lanceolate	1.21–2.87	0.27–0.57	0.10–0.35	0.27–0.37
Glinus lotoides	obovate or orbiculate to broadly spatulate obovate to oblong-spatulate orbiculate or more or less cuneate	0.90–2.33	0.46–0.80	0.34–0.47	0.04–0.44
Spinacia oleracea	hastate to ovate	1.22–1.63	0.08–0.39	0.17–0.25	0.37–0.63
	ovate to triangular-hastate oblong	1.81–2.21	0.45–0.55	0.27–0.33	0.27–0.33
Alternanthera paronychioides	oblanceolate or spatulate	2.83–3.46	0.62–0.76	0.28–0.34	0.09–0.11
	elliptic, ovate-rhombic, or oval	2.39–2.92	0.72–0.88	0.34–0.42	0.03–0.04
	elliptic, oval or obovate	1.45–2.57	0.40–0.69	0.17–0.38	0.22–0.32

ras. Some species only exist in particular regions, so parallel information is not guaranteed for each species.

In order to calculate the threshold for the integration, we selected a group of parallel descriptions from the dataset, which are not identical yet are still considered to be similar enough to be combined. The average distance of these parallel descriptions is used as the threshold, which turned out to be 0.98.

In Table 2, we list the original descriptions of several species with their integrated results. An overview of these parallel data is presented clearly. Some species' leaves, such as the first two, are described differently but all descriptions more or less agree with each other, therefore they are integrated into a single region with combined constraints on its four features. Here, the integration reduces the redundancies among the parallel information.

Other species, such as the last two, have quite different leaf shapes. These shapes are "dissimilar" enough to be kept as complementary information. If the species itself has wide variations, one author might not capture them all. Integration of parallel information makes the whole knowledge as nearly complete as possible. By comparing original descriptions and integrated ones, we can easily find some geographically-caused variations.

5 Results on Ranking of Responses to Queries

One of the advantages of putting NL information into a formal ontology is to make the knowledge in NL documents easier to access. After leaf shape information is represented formally, we can query species based on their leaf shapes. Similar to the method used in [4], firstly, the queried shape is represented by an OWL-Eu class Q. The shape reasoner interacts with FaCT-DG reasoner and returns a list of species, whose leaf shapes (in terms of the four features) either exactly match Q,are subsumed by Q,subsume Q (also called plugin matching), or intersect with Q.

Table 3. Query results for "lanceolate to elliptic" (partial)

Species	Leaf Shape Descriptions	Matching Type	Distance	Ranking
Comastoma muliense	lanceolate to elliptic	Exact	0.00	1
Polygonatum biflorum	narrowly lanceolate to broadly elliptic	Plugin	0.23	6
Hydrangea longifolia	lanceolate	Subsume	0.85	453
Rhodiola smithii	linear to oblong narrowly ovate to ovate-linear	Intersection	0.44	64

Some results for the query: "any possible species with lanceolate to elliptic leaves," is shown in Table 3. The matching type indicates the logic relations between the matched species and the query. Because our method uses the real semantics for querying, it can find some hidden results which are ignored by keyword matching, *e.g.*, the last species in Table 3. However, the problem is that it is not clear how well a result matches the query. The user has to go through the whole list and judge by himself.

Since our distance measure has been confirmed to be valid (see Section 4.1), we can use this distance as a criterion to quantify how well a logically matched species matches the query. A shorter distance means a better match. We sort the whole list based on the distance between each species' leaf shape and the queried one. Based on the matching ranks, as those in the last column of Table 3, the better matched results can be recovered easily.

We further enlarged our dataset from the eFloras, including 1154 species, some of which were described in more than one flora. Parallel descriptions were integrated first and then all queries are based on integrated knowledge. If one species has more than one shape region which matches the query, only the "best-match" (with the smallest distance to the query) is selected to join the ranking. We carried out 10 queries on basic terms and range phrases. Most queries finished in 1–2 seconds, the others took less than 5 seconds, on a 2G Hz Pentium 4 PC.

We compared our method with the keyword-based method over the 10 queries. For each query, results returned by both methods were merged into a single list, in the ascending order of their distances to the query. The ordered list was then divided into five groups, representing top 20% matched species, 20–40% matched ones, and so on. In each group, we counted the number of the species that the keyword-based method missed and that our method missed, respectively (see Figure 3 (a)). The numbers above each pair of columns is the mean similarity distance of that group. It shows that our method is able to find some well matched results (with small similarity distances) which are not matched by keyword search.

Due to the strictness of logic reasoning, our method failed to find some good results (judged by the expert). Therefore, we decreased the strictness level; if there are at least three features matched, the species is also returned if its distance to the query is less than the threshold which was used for integration. The performance was evaluated similarly, shown in Figure 3 (b). More hidden results were returned by our method while the quality (*i.e.*, mean distances) keeps stable.

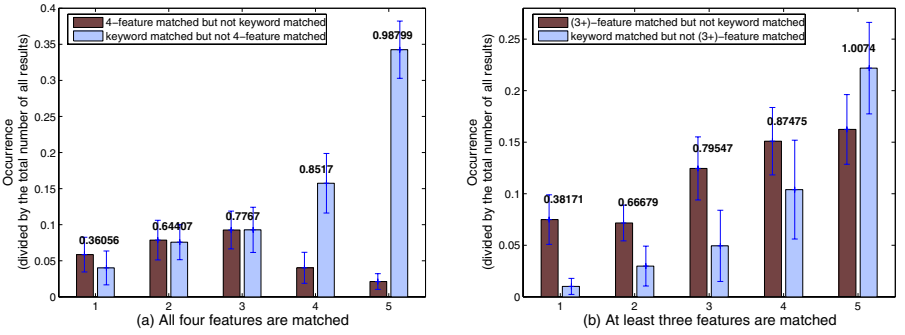


Fig. 3. Comparison of semantic-based query and keyword-based query

Table 4. Comparison between different levels of matching

Condition	Semantic matching		Keyword matching	
	Precision	Recall	Precision	Recall
4 features are perfectly matched	0.99888	0.55237	0.84474	0.65417
At least 3 features are matched	0.96642	0.72727	0.84430	0.65327

We use the standard precision/recall[12] to measure the performance of our method against keyword-based querying. From Table 4, we can see that when the strictness of matching criterion is loosened the precision decreases while the recall increases; this is a typical balancing problem.

In short, our approach outperforms the keyword-based method; this is because the former takes the real semantic of shape descriptions into account, while the latter simply checks the word matching of the descriptions.

6 Related Work

Sharing information among multiple sources occurs at many different levels. Access to a semantics is crucial for successful information integration and retrieval [1,3]. Instead of working on structured or semi-structured data, our work focuses mainly on integrating parallel NL information extracted from homogeneous monolingual (English) documents.

Many information integration systems have adopted ontologies as their working platform because of the various semantic expression of domain knowledge contained in ontologies [3,24,25] and powerful formal-logical reasoning tools supported by them [26,27,28]. Unfortunately, most systems stop at collecting and

[12] The precision indicates the proportion of answers in the returned list that were correct, while the recall is the proportion of correct answers in the whole data set that were found. Here, the correctness of a species is judged by whether the distance of its leaf shape description to the query is less than the integration threshold.

re-organising information from multiple sources instead of really integrating them based on their meanings.

The main obstacle for an ontology-based system to process NL documents is that the NL semantics is difficult to interpret. Many methods to capture and represent the semantics in NL have been tried, such as those multi-dimensional concept modelling including Osgood's semantic differential [29], lexical decomposition [30], etc. Using spatial or geometrical structures to model concepts has also been exploited in the cognitive sciences [31,14]. The limitations of their methods are either the dimensions are difficult to interpret or they are qualitative which prevents the semantics to be precisely captured.

It is not easy for a logic system to represent continuous ranges. OWL-Eu supports representing numerical ranges but still cannot express other ranges, *e.g.*, "ovate to elliptic". Using a semantic model to some extend helps the ontology system to represent such ranges. Furthermore, our work shows that datatype-enabled ontology reasoning can be very useful for real world applications.

Similarity measurement has been investigated in different knowledge representation systems and used in many applications [14,32], while similarity ranking is still one of the new ideas for current ontology techniques [33]. Traditionally, only subsumption checking is used to answer queries. Until recently, some other types of matching, such as intersection matching, are also considered for special cases [34]. However, there is little effort to integrate logic reasoning and similarity measuring. Such integration can determine how well results match the query and therefore can improve the usability of final results.

7 Conclusion

Ontology-based information integration in descriptive domains often comes to grief when comparison and integration have to be based on real semantics. Encouraged by our earlier work on processing parallel colour descriptions [4], we have applied the same methodology on leaf shape descriptions, where we introduced the notion of semantic distance to help parallel information integration and improve the usability of query results.

It turns out that the distances between shape descriptions are very hard to define. To solve the problem, we have derived a domain-dependent four feature leaf shape model. In our model, distances between the shapes are very well captured by the distances between the features, which has been evaluated by our domain expert. Besides the support of distance-based integration, our ontology-based approach (OA) outperforms the keyword-based approach (KA) because OA considers both the syntax and semantics of shape descriptions, while KA considers neither.

Most importantly, from the experiments in colour and leaf shape domain, we have learnt a set of more general methodological rules for processing parallel descriptive information in an ontology-based system. Key tasks we have identified include: (i) it is unlikely that a universal semantic model for all different domains exists, so for each domain, an appropriate (no need to be perfect)

model has to be chosen in order to get useful results; (ii) based on the semantic model, single terms have to be located, the effect of modifiers has to be defined and ranges have to be built properly; (iii) in order to integrate parallel information, a proper distance measurement is crucial to quantify the similarities among information from multiple sources; (iv) depending on the application, more expressive representation and additional reasoning may be necessary to solve real problems.

References

1. Stuckenschmidt, H., van Harmelen, F.: Information Sharing on the Semantic Web. Springer-Verlag (2004)
2. Ceusters, W., Smith, B., Fielding, J.M.: Linksuite: Formally robust ontology-based data and information integration. In: Proceedings of First International Workshop of Data Integration in the Life Sciences (DILS'04). Volume 2994 of Lecture Notes in Computer Science., Springer (2004) 124–139
3. Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Huebner, S.: Ontology-based integration of information - a survey of existing approaches. In: Proceedings of the IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA (2001) 108–117
4. Wang, S., Pan, J.Z.: Ontology-based representation and query colour descriptions from botanical documents. In: Proceedings of OTM Confederated International Conferences. Volume 3761 of Lecture Notes in Computer Science., Springer (2005) 1279–1295
5. Gielis, J.: A generic geometric transformation that unifies a wide range of natural and abstract shapes. American Journal of Botany **90** (2003) 333–338
6. Pan, J.Z., Horrocks, I.: OWL-Eu: Adding Customised Datatypes into OWL. In: Proceedings of Second European Semantic Web Conference (ESWC 2005). (2005) An extended and revised version is published in the Journal of Web Semantics, 4(1). 29-39.
7. Pan, J.Z.: Description Logics: Reasoning Support for the Semantic Web. PhD thesis, School of Computer Science, The University of Manchester (2004)
8. Pan, J.Z.: A Flexible Ontology Reasoning Architecture for the Semantic Web. In: IEEE Transactions on Knowledge and Data Engineering, Specail Issue on the Semantic Web. (2006) To appear.
9. Marr, D., Nishihara, H.: Representation and recognition of the spatial organization of three-dimensional shapes. In: Proceedings of the Royal Society B 200, London (1978) 269–294
10. Pentland, A.: Perceptual organization and the representation of natural form. Artificial Intelligence **28** (1986) 293–331
11. Thompson, D.: On growth and form. Cambridge University Press, London (1917)
12. Adams, D.C., Rohlf, F.J., Slice, D.E.: Geometric morphometrics: Ten years of progress following the "revolution". Italian Journal of Zoology **71** (2004) 5–16
13. Stearn, W.T.: Botanical Latin: history, grammar, syntax, terminology and vocabulary. David and Charles, Newton Abbot, England (1973)
14. Gärdenfors, P.: Conceptual Spaces: the geometry of thought. The MIT Press, Cambridge, Massachusetts (2000)
15. Clapham, A., Tutin, T., Moore., D.: Flora of the British Isles. Cambridge University Press (1987)

16. Stace, C.: New Flora of the British Isles. Cambridge University Press (1997)
17. Tutin, T.G., Heywood, V.H., Burges, N.A., Valentine, D.H., Moore(eds), D.M.: Flora Europaea. Cambridge University Press (1993)
18. Rose, F.: The Wild Flower Key: British Isles and North West Europe. Frederick Warne (1981)
19. Fernald, M.: Gray's Manual of Botany. American Book Company, New York (1950)
20. Dik, S.C.: Coordination: Its implications for the theory of general linguistics. North-Holland, Amsterdam (1968)
21. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., eds., L.A.S.: OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-ref/ (2004)
22. Carroll, J.J., Pan, J.Z.: XML Schema Datatypes in RDF and OWL. Technical report, W3C Semantic Web Best Practices and Development Group (2006) W3C Working Group Note, http://www.w3.org/TR/swbp-xsch-datatypes/.
23. Lydon, S.J., Wood, M.M., Huxley, R., Sutton, D.: Data patterns in multiple botanical descriptions: implications for automatic processing of legacy data. Systematics and Biodiversity (2003) 151–157
24. Goble, C., Stevens, R., Ng, G., Bechhofer, S., Paton, N., Baker, P., Peim, M., Brass, A.: Transparent access to multiple bioinformatics information sources. IBM Systems Journal Special issue on deep computing for the life sciences **40** (2001) 532 – 552
25. Williams, D., Poulovassilis, A.: Combining data integration with natural language technology for the semantic web. In: Proceedings of Workshop on Human Language Technology for the Semantic Web and Web Services, at ISWC'03. (2003)
26. Calvanese, D., Giuseppe, D.G., Lenzerini, M.: Description logics for information integration. In Kakas, A., Sadri, F., eds.: Computational Logic: Logic Programming and Beyond, Essays in Honour of Robert A. Kowalski. Volume 2408 of Lecture Notes in Computer Science. Springer (2002) 41–60
27. Maier, A., Schnurr, H.P., Sure, Y.: Ontology-based information integration in the automotive industry. In: Proceedings of the 2nd International Semantic Web Conference (ISWC2003), Sanibel Island, Florida, USA, Springer (2003) 897–912
28. Ferrucci, D., Lally, A.: UIMA: an architectural approach to unstructured information processing in the corporate research environment. Journal of Natural Language Engineering **10** (2004) 327–348
29. Osgood, C., Suci, G., Tannenbaum, P.: The Measurement of Meaning. University of Illinois Press, Urbana, IL (1957)
30. Dowty, D.R.: Word Meaning and Montague Grammar. D. Reidel Publishing Co., Dordrecht, Holland (1979)
31. Lakoff, G.: Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. University of Chicago Press (1987)
32. Schwering, A.: Hybrid models for semantics similarity measurement. In: Proceedings of OTM Confederated International Conferences. Volume 3761 of Lecture Notes in Computer Science., Springer (2005) 1449–1465
33. Anyanwu, K., Maduko, A., Sheth, A.P.: Semrank: ranking complex relationship search results on the semantic web. In Ellis, A., Hagino, T., eds.: WWW, ACM (2005) 117–127
34. Li, L., Horrocks, I.: A Software Framework For Matchmaking Based on Semantic Web Technology. In: Proceedings of the Twelfth International World Wide Web Conference (WWW 2003), ACM (2003) 331–339